PRO GRADU -TUTKIELMA

Mikko Korhonen

Katoanalyysi nuorten terveystutkimuksessa

TAMPEREEN YLIOPISTO
Luonnontieteiden yksikkö
Tilastotiede
Huhtikuu 2017

Tampereen yliopisto

Luonnontieteiden yksikkö

KORHONEN, MIKKO: Katoanalyysi nuorten terveystutkimuksessa

Pro Gradu -tutkielma, 32 s., 0 liites.

Tilastotiede

Huhtikuu 2017

Tiivistelmä

Tutkielmassa tarkastellaan puuttuvaa tietoa Pirkanmaan sairaanhoitopiirin tutkimuksessa, joka käsittelee nuorten mielenterveyteen liittyviä ominaisuuksia. Analysoitavana aineistona on Pirkanmaan sairaanhoitopiirin teettämät kyselyt tamperelaisille ja vantaalaisille nuorille vuosina 2002, 2004 ja 2010. Tutkimus on kolmiosainen, ja siihen vastaaminen on vapaaehtoista. Kyselyn vastausprosentti oli 2004 63,1 % ja 2010 39,2 %. Tutkielmassa tavoitteena on löytää millaiset ominaisuudet riippuvat siitä, jättivätkö haastateltavat myöhemmin vastaamatta yhteen tai useampaan kyselyyn. Tällöin käytetään analyysissä haastateltavan aiempia vastauksia. Vastaajat jakautuvat kolmeen ryhmään sen mukaan, jättivätkö he vastaamatta yhteen tai useampaan kyselyyn.

Aineiston analysointi aloitettiin tarkastelemalla kaikkia kolmea vastaajaryhmää yhtäaikaisesti. Tällöin käytettiin ainoastaan ensimmäisen kyselyn vastauksia, kun tutkittiin erosavatko ryhmien vastaukset toisistaan. Kruskal-Wallisin testillä havaittiin useiden muuttujien jakaumien eroavan ryhmien välillä tilastollisesti merkittävän paljon. Merkitseviin muuttujiin kuuluivat muun muassa päihteiden käyttöä kuvaavat muuttujat humalahakuinen juominen ja hasiksen käyttö, sekä nuoren terveyden tilaa kuvaavat rikekäyttäytyminen ja sosiaalinen tuki. Testaamisen lisäksi tarkasteltiin tilastollisia tunnuslukuja, joiden perusteella vaikutti, että useissa ryhmien välillä eroavissa muuttujissa oli trendiä, kun ohitettujen kyselyiden määrä kasvoi.

Analysoinnin toisessa osassa tutkittiin vastaajaryhmän riippuvuutta haastateltavien vastausten kanssa, kun vertailtiin lopettaneiden vastaajaryhmiä erikseen kaikkiin vastanneiden ryhmän kanssa. Tässä käytettiin etenevää valintaa lineaariselle ja neliölliselle luokitteluanalyysille sekä ristiinvalidointia. Tulosten tarkastelussa hyödynnettiin merkkitestiä, jolla selvitettiin, voitiinko aineiston muuttujia pitää tilastollisesti merkitsevästi parempina kuin generoitua satunnaismuuttujaa. Tällöin tutkittiin, sijoittuiko muuttuja generoitua muuttujaa paremmin etenevissä valinnoissa. Lisäksi tarkasteltiin, miten hyvin malleilla voitiin ennustaa haastateltavan vastaajaryhmä.

Tarkasteltaessa ensimmäisen vaiheen jälkeen lopettaneita kaikkiin vastanneiden kanssa havaittiin useiden muuttujien olevan satunnaisuutta parempia. Vähintään toisessa luokitteluanalyysissä tärkeitä muuttujia olivat muun muassa humalahakuinen juominen, huumausaineiden käyttö viimeisen kuukauden aikana ja sosiaalinen tuki. Lisäksi mallien ennustustarkkuutta tarkastellessa ryhmiteltiin mallit sen mukaan, montako selittäjää niissä oli. Tällöin havaittiin mallien ennustavan satunnaisuutta paremmin kaikilla muuttujien lukumäärillä, kun tutkittiin ryhmien keskimääräisiä prosentteja. Lineaarisen ja neliöllisen luokitteluanalyysin mallien välillä ei pääasiassa ollut tilastollisesti merkitseviä eroja, vaikka lineaarinen luokitteluanalyysi antoikin otoksessa hieman korkeamman prosentin.

Tutkittaessa toisessa vaiheessa lopettaneita kaikkiin vastanneiden kanssa havaittiin jossain määrin vastaavia riippuvuuksia kuin edellisessä analyysissä. Merkittäviä muuttujia olivat muun muassa humalahakuinen juominen, aggressiivinen käytös, huumausaineiden käyttö viimeisen kuukauden aikana ja toimettomuus. Ensimmäinen muuttuja oli merkittävä sekä ensimmäisessä että toisessa kyse-

lyssä, ja muut mainitut olivat toisesta kyselystä. Tässäkin tarkastelussa mallit antoivat keskimäärin satunnaisuutta parempia ennusteita. Lineaarisen ja neliöllisen luokitinten ennustustarkkuudet eivät eronneet tilastollisesti merkitsevästi toisistaan.

Analyysin lopulla tarkasteltiin haastateltavien vastausten vaihtelua pääkomponenttianalyysillä ja Sammon kartalla. Tällöin tutkittiin ensin ensimmäisen kyselyn vastauksia ja sitten toisen. Kummassakaan tarkastelussa ei havaittu vastaajaryhmien välillä selvää erottuvuutta. Sen sijaan humalahakuinen juominen jakautui molempien kyselyiden kaksiulotteisessa Sammon kartassa kahdeksi ryhmäksi.

Asiasanat: Kruskal-Wallisin testi, neliöllinen luokitteluanalyysi, lineaarinen luokitteluanalyysi, merkkitesti, Sammon kartta

Sisällysluettelo

1 Johdanto

Tässä tutkielmassa tarkastellaan nuorten terveyskyselyssä vastanneita ja vastaamisen keskeyttäneitä. Tavoitteena on löytää yhteyksiä vastaamiskäyttäytymisen ja vastausten välillä. Alaluvussa 1.1 tarkastellaan tutkielman tarkoitusta ja alaluvussa 1.2 tutkielman taustaa.

1.1 Tutkielman tarkoitus

Tutkielmassa analysoidaan Pirkanmaan sairaanhoitopiirin teettämää ”Nuorten terveys”-tutkimusta. Tutkimukseen vastasivat tamperelaiset ja vantaalaiset nuoret. Kysely oli kolmiosainen, ja se toteutettiin vuosina 2002, 2004 ja 2010. Vastaajat olivat keskimääräisesti iältään 15,5, 17,6 ja 22,8 vuotta. Kyselyyn vastaaminen oli vapaaehtoista.

Tässä tutkimuksessa keskitytään tarkastelemaan niitä tekijöitä, joilla on yhteyttä sen kanssa, lopettiko nuori kyselyihin vastaamisen. Vastaajat on jaettu kolmeen ryhmään sen mukaan, miten he vastasivat tutkimukseen. Kaikkiin kolmeen kyselyyn vastanneet ovat omana ryhmänään, vain kahteen ensimmäiseen vastanneet ovat oma ryhmänsä, ja loput muodostavat kolmannen ryhmän, olivat he vastanneet kolmanteen kyselyyn tai eivät. Ryhmien osuudet ovat samassa järjestyksessä 31,9 %, 31,2 % ja 36,9 %.

Tutkielmassa tutkitaan ensin kolmen ryhmän samankaltaisuutta testaamalla, jakautuvatko kunkin muuttujan arvot samoin ryhmissä. Lisäksi tarkastellaan erikseen lopettaneiden ryhmiä, kun heitä verrataan kaikkiin vastanneisiin. Tällöin tutkitaan, mitkä tekijät ennustavat parhaiten lopettiko nuori vastaamisen. Selittävinä muuttujina käytetään nuoren terveyttä sekä perhetilannetta kuvaavia muuttujia. Selittävien muuttujien vaihtelua tarkastellaan myös ilman selitettävää muuttujaa.

1.2 Tutkielman tausta

Tässä tutkielmassa on tarkoitus selvittää, millä tekijöillä on riippuvuutta sen kanssa, vastaako haastateltava kaikkiin kyselyihin. Tavoitteena on kartoittaa, millä tekijöillä on riippuvuuksia eri ryhmis-

sä ja minkä suuntaisia nämä ovat. Käsiteltävä tutkimus on pitkittäisdata, josta on mahdollista löytää vastaamisen keskeyttäneiden aiemman vastaukset.

Tiedonkeruussa merkittävä ongelma on puuttuva tieto, erityisesti mikäli kyseessä on puuttuva havainto. Tällöin otoksen vastaukset eivät kuvaa koko perusjoukkoa, vaan jotain tämän osajoukkoa, jolloin tutkimuksessa saattaa esiintyä harhaa. Erityisesti mikäli analyyseissa havaitaan jonkin muuttujan riippuvan voimakkaasti vastaamiskäyttäytymisen kanssa, voidaan tämän perusteella saada arviota siitä, ovatko muuttujan jotkin arvot yli- tai aliedustettuina jossakin ryhmässä. Tutkiessa psykologian alan aineistoa voitaisiin ajatella havaittavan yhteyksiä vastausten ja keskeyttämisen välillä, mikäli vastauskäyttäytyminen riippuu tutkittavista luonteenpiirteistä. Suominen, Koskenvuo ja Sillanmäki (2012) havaitsivat aikuisille suunnatussa kyselyssä vastaamiskäyttäytymisen riippuvan selvästi vastaajan terveyttä koskevista ominaisuuksista.

Fröjd, Kaitala-Heino ja Marttula (2010) havaitsivat tutkimuksessaan vastaajaryhmän muun muassa riippuvan tamperelaisten ja vantaalaisten nuorten keskuudessa YSR-muuttujasta (ks. luku 2.2), ja tamperelaisten nuorten vastausryhmä riippui masennuksesta. Kyseiset päätelmät oli tehty χ^2-riippumattomuustestin sekä Fisherin tarkan testin perusteella. Analyysissä käsiteltiin kaksiosaista tutkimusta, jossa vaiheiden väli oli kaksi vuotta, ja vastaajat olivat keskimääräisesti iältään 15–16 vuotta ensimmäisessä vaiheessa ja 17–18 vuotta jälkimmäisessä. Tässä tutkielmassa tarkastellaan muiden muuttujien ohella, onko vastaavat riippuvuudet havaittavissa tutkittavassa aineistossa, kun kyselyjä on kolme. Lisäksi tässä ei rajoituta ainoastaan parametrisiin testeihin vastaajaryhmien vertailussa.

2 Aineiston kuvailu

Tarkastellaan seuraavaksi analyysissä käytettäviä muuttujia. Ensin alaluvussa 2.1 kuvaillaan kategorisia muuttujia, joita on sekä luokittelu- että järjestysasteikollisia. Tämän jälkeen alaluvussa 2.2 tarkastellaan numeerisia muuttujia.

2.1 Kategoriset muuttujat

Alkuperäisessä aineistossa oli 1186 muuttujaa ja 3278 havaintoa. Käytetyssä aineistossa on 26 tutkittavaa muuttujaa, osa näistä on saatu alkuperäisistä muuttujista summaamalla, lisäksi suurin osa muuttujista on rajattu pois. Tutkittavissa muuttujissa on puuttuvaa tietoa noin 1,64 %, kun ei huomioida kyselyitä, joihin vastaaja ei ole vastannut lainkaan.

Vastaajat on jaettu kolmeen ryhmään sen mukaan, miten he vastasivat tutkimukseen. Kaikkiin kolmeen kyselyyn vastanneet ovat omana ryhmänään, vain kahteen ensimmäiseen vastanneet ovat oma ryhmänsä, ja loput muodostavat kolmannen ryhmän, olivat he vastanneet kolmanteen kyselyyn tai eivät. Ryhmien osuudet ovat samassa järjestyksessä 31,9 %, 31,2 % ja 36,9 %. Tässä tutkielmassa *vastaajaryhmällä* viitataan siihen, mihin edellä mainituista ryhmistä vastaaja kuului.

Ensimmäisenä tutkittavana muuttujana on *koettu terveys*, joka kuvaa sitä, miten hyväksi nuori kokee oman senhetkisen terveydentilansa. Vastausvaihtoehtoina olivat erittäin hyvä, melko hyvä, keskinkertainen, melko huono ja erittäin huono. Alaluvussa 4.2 kaksi edeltävää on yhdistetty yhdeksi ryhmäksi ja kolme seuraavaa toiseksi. Kysymys on toistettu jokaisessa kolmessa kyselyssä. Ensimmäisessä kyselyssä 17 % vastaajista koki terveytensä keskinkertaiseksi tai heikommaksi, ja toisessa kyselyssä osuus oli 19 %.

Humalahakuista juomista mitataan sillä, kuinka usein nuori juo itsensä tosi humalaan. Tässä vaihtoehtoina ovat kerran viikossa tai useammin, noin 1-2 kertaa kuukaudessa, harvemmin tai ei koskaan. Alaluvussa 4.2 kolme ensimmäistä vaihtoehtoa on yhdistetty yhdeksi ryhmäksi, ja viimeinen ryhmä on pysynyt ennallaan. 2002 kyselyssä 57 % kertoi juovansa itsensä humalaan vähintään joskus, ja 2004 vastaava prosentti oli 72.

Päihteiden sekakäyttö kuvaa miten usein nuori on käyttänyt alkoholia sekä lääkkeitä samanaikaisesti. Vastausvaihtoehtoina ovat ei kertaakaan, kerran, 2-4 kertaa tai useammin. Vuoden 2004 kyselyssä aikavälinä oli viimeinen kaksi vuotta. Alaluvussa 4.2 ensimmäinen ryhmä on pidetty ennallaan ja kolme seuraavaa on yhdistetty yhdeksi ryhmäksi. Vuonna 2002 10 % oli käyttänyt alkoholia

ja lääkkeitä samanaikaisesti. 2004 osuus oli pudonnut 7 prosenttiin.

Voimakkaiden huumeiden käyttö kuvaa kuinka monta kertaa nuori on käyttänyt voimakkaaksi luokiteltavaa huumausainetta, toisessa kyselyssä aikavälinä oli viimeinen kaksi vuotta. Tässä vaihtoehdot ovat samat kuin päihteiden sekakäytössä ja alaluvussa 4.2 ryhmät on yhdistetty kahdeksi vastaavasti kuin edellä. Kahdessa ensimmäisessä kyselyssä noin 3 % oli käyttänyt kovia huumeita.

Viimeaikainen huumeiden käyttö kuvaa kuinka monta kertaa nuori on käyttänyt huumausaineita viimeisen 30 päivän aikana. Tässä ryhmät ovat samat kuin edellä. 2002 kyselyssä 20 % oli käyttänyt huumeita viimeisen kuukauden aikana, ja 2004 osuus oli pudonnut 13 prosenttiin.

Viimeinen päihteidenkäyttöä kuvaava muuttuja on *hasiksen käyttö*. Muuttujassa ensimmäisen ryhmän muodostavat ne, jotka eivät ole ikinä kokeilleet hasista, ja toisen ryhmän ne, jotka ovat kokeilleet ainakin kerran. 2002 hasista oli käyttänyt 12,9 % ja kahden vuoden kuluttua vastaava osuus oli 15,8 %.

Muuttuja *perheen taloudelliset vaikeudet* kuvaa onko nuori kokenut perheellä olevan taloudellisia ongelmia. Alkuperäisessä luokittelussa kysyttiin myös nuoren suhtautumista mahdollisiin vaikeuksiin, mutta luokat on yhdistetty siten, että onko nuori kokenut perheellä olevan vaikeuksia. Vastaajista 19 % oli kokenut taloudellisia vaikeuksia 2002 kyselyssä, ja osuus oli toisessa kyselyssä 18 %.

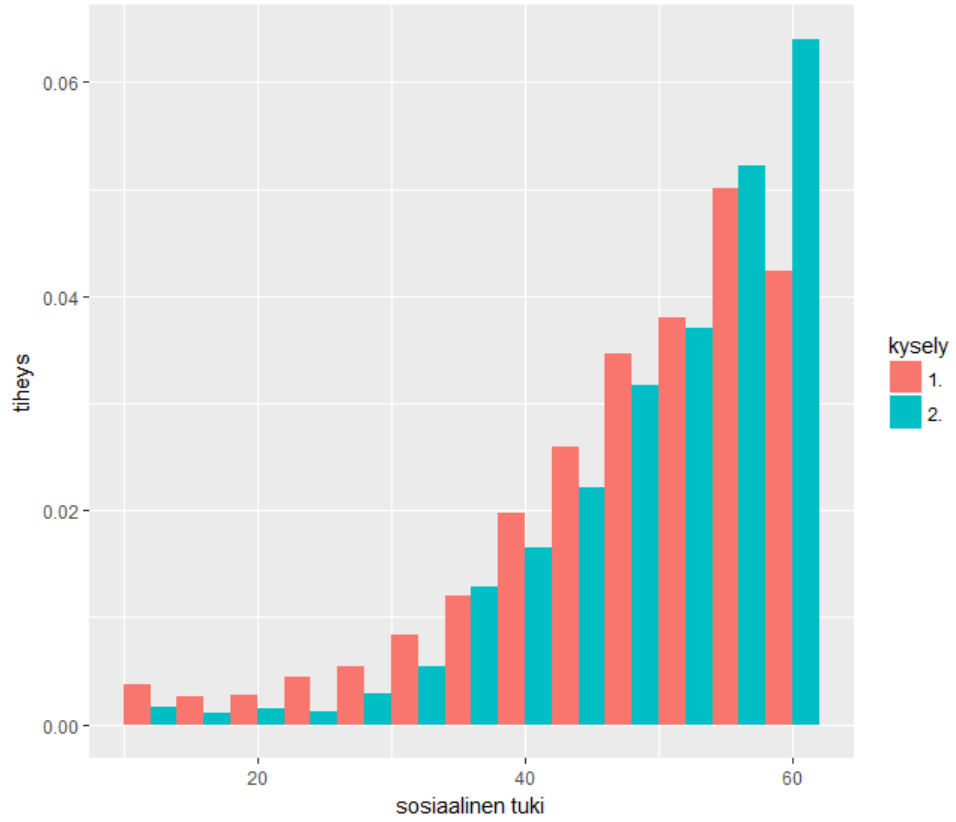
Nuoren syrjäytymistä kuvataan muuttujalla tämän pääasiallisesta toimesta. *Toimettomuutta* kuvaa vaihtoehto ”kotona muuten vaan päätoimisesti”, muita vaihtoehtoja ovat muun muassa työskentely, opiskelu, vanhempainvapaa, armeija, siviilipalvelus ja kuntoutus joko päätoimisesti tai osapäiväisesti. Kysymystä ei esitetty ensimmäisessä kyselyssä, toisessa kyselyssä 3 % kertoi olevansa päätoimisesti kotona.

2.2 Numeeriset muuttujat

Nuoren saamaa sosiaalista tukea on mitattu muuttujalla *PSSS* (Perceived Social Support Scale -Revised). Tämä on numeerinen muuttuja, joka on laskettu summaamalla 12 sosiaalista tukea kuvaavaa muuttujaa. Kysymykset kuvaavat miten hyvin nuori saa läheisiltään tukea, kun vastaukset on pisteytetty yhdestä viiteen, kun pienet arvot kuvaavat vähäistä tukea. Vastaavasti tulkitaan PSSS-muuttujaa, joka saa arvoja väliltä 12–60. (Blumenthal et al. 1987)

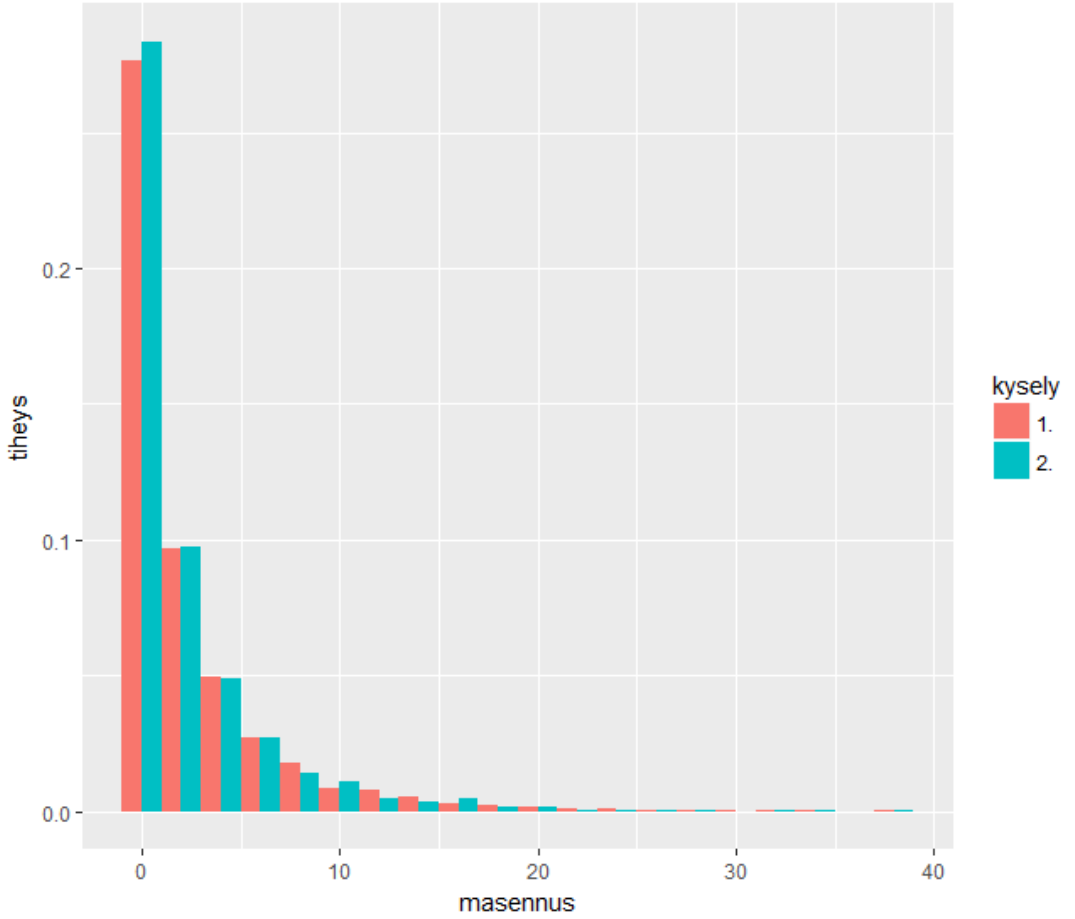
Pistemäärien jakautumista kahdessa ensimmäisessä kyselyssä on kuvattu kuviossa 2.1. Kuvion perusteella vaikuttaisi, että ensimmäiseen ja toiseen kyselyyn vastanneiden välillä sosiaalisen tuen määrä ei eroa merkittävästi toisistaan lukuun ottamatta hyvin suuria arvoja, jotka ovat toisessa kyse-

lyssä hieman yleisempiä.



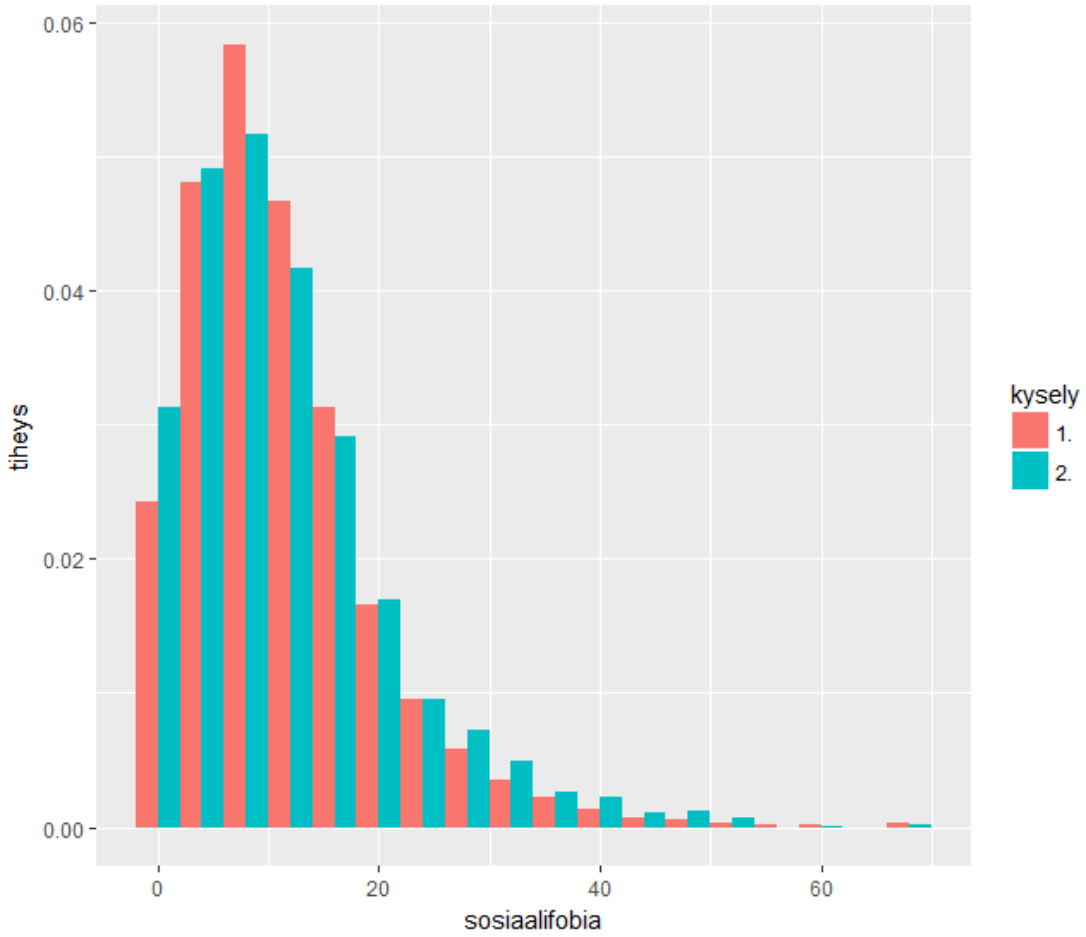
Kuvio 2.1. Histogrammit sosiaalista tukea kuvaavan muuttujan PSSS jakautumisesta kyselyissä 2002 (punainen) ja 2004 (sininen). Y-akselilla ”tiheys” kuvaa pistemäärän yleisyyttä, nämä summautuvat molemmissa kyselyissä ykköseksi, kun huomioidaan pylvään leveys.

Ensimmäinen tutkittava sairaus on *masennus*. Luokittelu perustuu RBDI-kyselyyn, jossa haastateltava vastaa kolmeentoista mielialaa koskevaan kysymykseen (Raitasalo 2007, s. 22–24). Vastaukset on pisteytetty kokonaisluvuilla väliltä 0–3, joten yhteispistemäärä on väliltä 0–39. Pistemäärän kasvaessa masennusoirehdinta lisääntyy. Mikäli vastaaja vastasi vähintään 10:en kyselyn kysymykseen ja jätti jonkin kysymyksen tyhjäksi, on tyhjät korvattu vastaajan omalla keskiarvolla muista masennuskysymyksistä. Masennuksen jakautumista kahdessa ensimmäisessä kyselyssä on kuvattu kuviossa 2.2. Kuvion perusteella masennuksen jakautuneisuudessa ei näy selviä eroja kyselyjä vertaillessa.



Kuvio 2.2. Histogrammikuviot masennuksen jakautumisesta kyselyissä 2002 (punainen) ja 2004 (sininen). Y-akselilla ”tiheys” kuvaa pistemäärän yleisyyttä, arvot summautuvat molemmissa kyselyissä ykköseksi, kun huomioidaan pylvään leveys, joka on tässä tapauksessa 2.

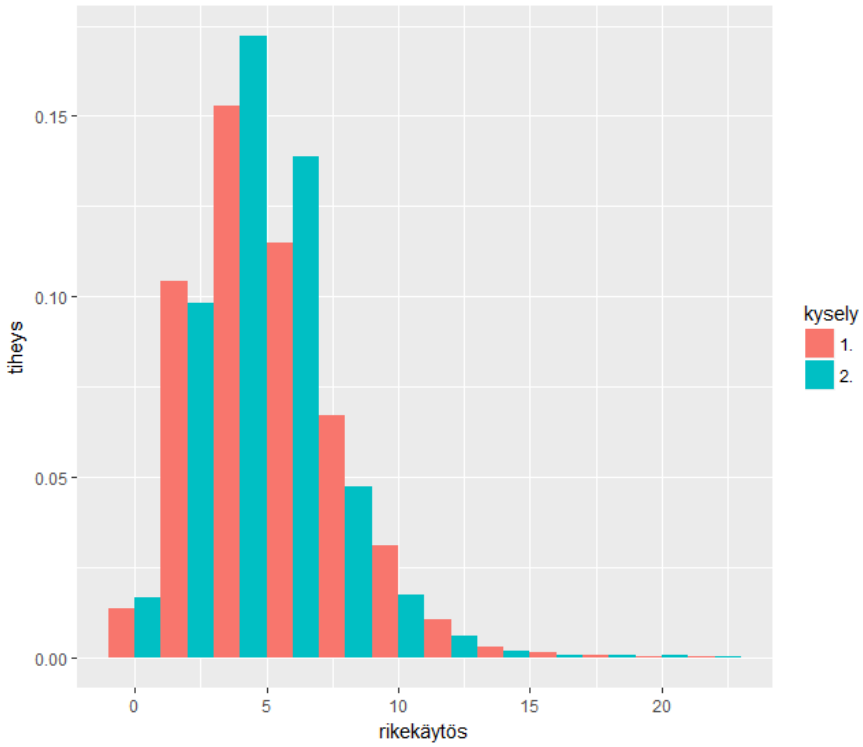
Sosiaalinen ahdistuneisuushäiriö eli *sosiaalifobia* perustuu kyselytutkimuksessa tehtyyn The Social Phobia Inventory- eli SPIN-kyselyyn, jossa on 17 väittämää (Churchill et al. 2000, s. 176, 379–386). Haastateltava valitsee 0–4 pisteytetyistä vastausvaihtoehdoista yhden sen mukaan, miten hyvin väite kuvaa häntä. Täten SPIN-muuttujan yhteispistemäärä on väliltä 0–68. Sosiaalisen ahdistushäiriön oireet lisääntyvät pistemäärän kasvaessa. Mikäli vastaaja on vastannut vähintään 13 kysymykseen, on mahdolliset puuttuvat arvot korvattu tämän vastausten keskiarvolla. Sosiaalifobian yleisyyttä kyselyissä 2002 ja 2004 on kuvattu kuviossa 2.3. Kyselyiden välillä ei näy selviä eroja sosiaalifobian yleisyydessä.



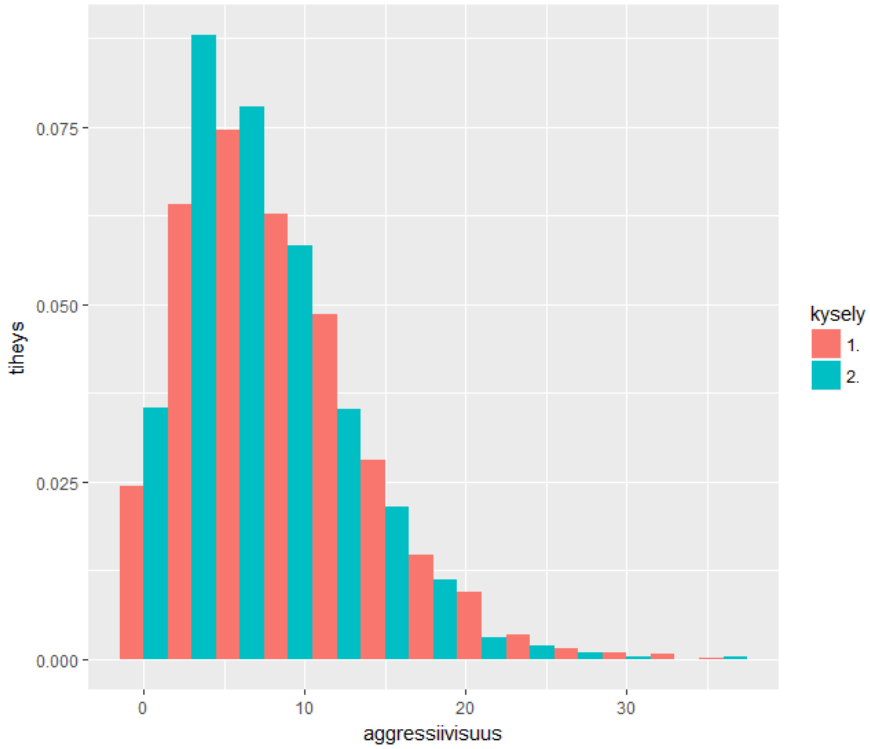
Kuvio 2.3. Histogrammit sosiaalifobian jakautumisesta kyselyissä 2002 punaisella ja 2004 sinisellä. Y-akselilla ”tiheys” kuvaa pistemäärän yleisyyttä, kun tiheyden summautuvat molemmissa kyselyissä ykköseksi, kun huomioidaan pylvään leveys.

Käytöshäiriökysymykset pohjautuvat YSR-kyselyyn, jossa on 29 erilaista häiriökäyttäytymistä kuvaavaa kysymystä (Achenbach & Rescorla 2001). Haastateltavalla on kolme vastausvaihtoehtoa, jotka on pisteytetty sen mukaan, miten usein häiriökäyttäytymistä esiintyy. 11 ensimmäistä kysymystä kuvaa *rikekäyttäytymistä*, kuten epäsosiaalista ja rikollista käytöstä. Jos vastaaja on vastannut vähintään 9:ään näistä kysymyksistä, on puuttuvat korvattu vastaajan keskiarvolla muista osion kysymyksistä. Loput kahdeksantoista muuttujaa kuvaavat *aggressiivista käyttäytymistä*, joissa puuttuvat arvot on korvattu muiden osion vastausten keskiarvolla, mikäli vastaaja on vastannut vähintään 14 kysymykseen.

Rikekäyttäytymisen jakautuneisuutta kahdessa ensimmäisessä kyselyssä on havainnollistettu kuviossa 2.4. Kuvion perusteella vaikuttaisi, että ensimmäisessä kyselyssä rikekäyttäytyminen oli lievästi yleisempää. Aggressiivisuutta on havainnollistettu vastaavasti kuviossa 2.5. Tässä näyttäisi olevan samansuuntaista eroa ryhmien välillä, tosin vielä selvempää.



Kuvio 2.4. Rikekäyttäytymisen histogrammit ensimmäisissä kyselyissä 2002 (punainen) ja 2004 (sininen). Tiheys kuvaa kyseisen arvojoukon yleisyyttä kyselyssä, ja arvot summautuvat molemmissa kyselyissä 1:ksi, kun huomioidaan pylvään leveys.



Kuvio 2.5. Aggressiivisen käyttäytymisen histogrammit kyselyissä 2002 (punainen) ja 2004 (sininen). Tiheys kuvaa kyseisen arvojoukon yleisyyttä kyselyssä, ja se summautuu molemmissa kyselyissä ykköseksi, kun huomioidaan pylvään leveys.

3 Aineiston analysointimenetelmät

Tässä luvussa käydään läpi aineiston analysoinnissa käytetyt menetelmät. Kruskal-Wallisin testiä käytetään alaluvussa 4.1 testattaessa vastaajaryhmien vastausten jakaumien yhtäsuuruutta. Alalukujen 3.2–3.6 luokitteluanalyyseihin ja etenevään valintaan liittyviä menetelmiä sovelletaan aineiston analysoinnin alaluvuissa 4.2.1–4.2.2 luokitteluanalyyseissä sekä niiden tulosten analysoinnissa. Alaluvun 3.7 datan ulottuvuuksien vähentämismenetelmiä sovelletaan alaluvussa 4.2.3 havainnollistaessa aineiston vaihtelua Sammon kartalla.

3.1 Kruskal-Wallisin testi

Tavoitteena on tutkia sitä, ovatko haastateltujen vastaukset jakautuneet samoin eri vastaajaryhmissä. Valitaan ensin yksi aineiston muuttujista ja tutkitaan tämän jakautumista. Nyt tutkittavana on vähintään ordinaalinen muuttuja Y, jota selitetään kategorisella muuttujalla X, joka on haastateltavan vastaajaryhmä. Y kuvaa haastateltavan vastausta tutkittavaan kysymykseen. Tässä selitettävän oletetaan olevan luonteeltaan suhdeasteikollinen, vaikka data saattaa olla muodostettu yksinkertaistetulla asteikolla. Muuttujalla X on j eri mahdollista arvoa. Tutkimusongelmana on, jakautuuko Y samoin X:n muodostamissa osajoukoissa.

Tässä käytetään Kruskal-Wallisin testiä, jossa nollahypoteesina on, että Y jakautuu samoin X:n muodostamissa ehdollisissa osapopulaatioissa. Vaihtoehtoinen hypoteesi on, että ainakin jokin jakauma poikkeaa muista. Testissä havaintoaineisto järjestetään muuttujan Y mukaiseen suuruusjärjestykseen. Tämän jälkeen havainnoille voidaan laskea edelliseen perustuvat sijaluvut. Näistä laskettava testisuure H noudattaa nollahypoteesin ollessa tosi likimain χ^2-jakaumaa vapausasteilla j–1, kun otoskoko on riittävän suuri. Tällöin voidaan laskea jakauman kertymäfunktion approksimoinnin avulla p-arvo, jonka ollessa riittävän suuri nollahypoteesi jakaumien yhtäsuuruudesta voidaan hyväksyä. Testauksessa ei ole tarpeen olettaa selitettävän muuttujan olevan normaalisti jakautunut. (Hollander & Wolfe 1973.)

3.2 Lineaarinen erotteluanalyysi

Oletetaan että tutkittavana on selittävien muuttujien matriisi $\mathbf{X}=(X_1, X_2,..., X_m)$ sekä selitettävä vas-

temuuttuja Y, joka koostuu n:stä havainnosta, ja jonka soluilla on kaksi mahdollista arvoa: 0 ja 1. Tällöin matriisissa **X** on n havaintoa, jotka ovat m-ulotteisia vektoreita. Havainnot oletetaan toisistaan riippumattomiksi. Havainnot voidaan jakaa vastemuuttujan arvojen perusteella populaatioihin α_0 ja α_1. Matriisin **X** muuttujat ovat joko mitta-asteikollisia tai binäärisesti koodattuja. Lisäksi selittävät muuttujat on skaalattu siten, että jokaisen muuttujan keskiarvo on 0 ja otosvarianssi 1. Oletetaan populaation α_i havaintojen noudattavan normaalijakaumaa odotusarvovektorilla $\underline{\mu}_i$ ja kovarianssimatriisilla **Σ**, kun i=0,1. Tällöin valitun vektorin $\underline{x}$ tiheysfunktio on kuten kaavassa (3.1), kun vektorin tiedetään kuuluvan populaatioon α_i:

$$f(\underline{x}\ |\alpha_i)=\frac{1}{(2\pi)^{m/2}|\mathbf{\Sigma}|^{1/2}}\exp\left(-\frac{1}{2}(\underline{x}\ -\underline{\mu}_i)^T|\mathbf{\Sigma}|^{-1}(\underline{x}\ -\underline{\mu}_i)\right). \tag{3.1}$$

Lineaarisessa erotteluanalyysissä pyritään löytämään selittäville muuttujille sellainen lineaarikombinaatio, että se erottelee selitettävän muuttujan määräämät ryhmät mahdollisimman hyvin. Tällöin oletetaan myös, että kovarianssimatriisit ovat samoja eri populaatioissa. Olkoon $\underline{a}=(a_1, a_2,..., a_m)$ skalaarikerroinvektori, jonka varianssi on vakioitu. Tällöin tutkittavana on lauseke (3.2):

$$y=\underline{a}^T\boldsymbol{X}=\sum_{i=1}^{m}a_iX_i=a_1X_i+...a_mX_m\ ,\ \text{missä}\ \ Var(\underline{a}^T\boldsymbol{X})=\underline{a}^TVar(\boldsymbol{X})\underline{a}=\underline{a}^T\mathbf{\Sigma}\underline{a}\ \ \text{vakioitu.} \tag{3.2}$$

(Johnson & Wichern 2007.)

Tässä tutkielmassa erotteluanalyysiä ei varsinaisesti käytetä aineiston analyysiin, vaan ainoastaan tähän perustuvaa luokitteluanalyysiä. Täten tässä ei käydä läpi erotteluanalyysistä muuta kuin menetelmän periaate.

3.3 Neliöllinen erotteluanalyysi

Edellisessä alaluvussa yhtenä oletuksena lineaariselle erotteluanalyysille oli, että ryhmien väliset kovarianssimatriisit ovat yhtä suuria. Neliöllisessä erotteluanalyysissä voidaan luopua tästä oletuksesta, joten havainnot oletetaan ainoastaan normaalisti jakautuneiksi molemmissa populaatioissa sekä toisistaan riippumattomiksi. Tässäkin tarkastelussa pyritään löytämään menetelmä, joka erottelee eri populaatioiden havainnot niin hyvin kuin mahdollista. Kun pyritään löytämään sääntö havainnon $\underline{x}$ luokittelemiseen, tällöin tutkitaan lauseketta (3.3):

(3.3) $$\underline{x}^T \boldsymbol{A}\, \underline{x} + \underline{b}^T \underline{x} + \underline{c}.$$

Tässä **A** on tuntematon reaalimatriisi, ja $\underline{b}$ ja $\underline{c}$ tuntemattomia reaalivektoreita. Tavoitteena olisi siis löytää sellainen sääntö, että edellä mainitun lausekkeen arvot eroaisivat vastemuuttujan määräämien ryhmien välillä mahdollisimman paljon. (Johnson & Wichern 2007.)

Havaintojen luokittelussa tärkeämpää kuitenkin on tarkastella havaintojen tiheysfunktiota, joka oletetaan normaalijakaumaksi. Kun parametreja merkitään samoin kuin edellisessä alaluvussa sekä populaation α_i kovarianssimatriisi on $\boldsymbol{\Sigma_i}$, ovat populaation α_i havainnot jakautuneet kuten kaavassa (3.4):

(3.4) $$f(\underline{x} \mid \alpha_i) = \frac{1}{(2\pi)^{m/2} |\boldsymbol{\Sigma_i}|^{1/2}} \exp\left(-\frac{1}{2} (\underline{x} - \underline{\mu_i})^T |\boldsymbol{\Sigma_i}|^{-1} (\underline{x} - \underline{\mu_i}) \right).$$

3.4 Luokitteluanalyysi

Erotteluanalyysiä sovelletaan luokitteluanalyysissä, jossa ennustetaan jo havaituille vektoreille näiden luokat, jotka ovat tunnettuja. Tällöin luokittelun tarkkuutta tutkimalla voidaan havaita, millä muuttujilla on riippuvuuksia ryhmien kanssa. Jos tutkittava havainto on vektori $\underline{y}$, niin populaation α_i posterioritodennäköisyyden tiheysfunktio $f_p(\alpha_i \mid \underline{y})$ havainnolle voidaan määrittää Bayesin kaavalla (Lindgren 1976). Prioritodennäköisyytenä on populaation todennäköisyys, joka on Bernoullinjakautunut. Merkitään todennäköisyysfunktiota funktiolla p. Koska todennäköisyys- ja tiheysfunktiot ovat positiivisia, voidaan Bayesin kaavalla määrittää myös käytännöllisempi logaritmoitu posteriorifunktio $f_p(\alpha_i \mid \underline{y})$, joka on esitetty kaavassa (3.5):

(3.5) $$f_p(\alpha_i \mid \underline{y}) = \frac{f(\underline{y} \mid \alpha_i)\, p(\alpha_i)}{f_t(\underline{y})} \Rightarrow \log f_p(\alpha_i \mid \underline{y}) = \log f(\underline{y} \mid \alpha_i) + \log p(\alpha_i) - \log f_t(\underline{y}).$$

Edellisessä kaavassa funktio f kuvaa erotteluanalyysissä käytettyä tiheysfunktiota, ja f_t kuvaa kokonaistodennäköisyyden tiheysfunktiota. Nyt voidaan pitää mielekkäänä luokitella uusi havainto siihen populaatioon, jossa populaation uskottavuus on suurempi, eli kumman posteriori on suurempi. Koska luonnollinen logaritmifunktio on aidosti kasvava, on logaritmoidun posteriorifunktion tutkiminen yhtäpitävää edellisen lausekkeen kanssa. Kun käytetään neliöllistä luokittelua, on uskotta-

vuusfunktio samaa muotoa kuin kaavassa (3.4), ja tarkasteltava logaritmoitu posteriorifunktio on kuten kaavassa (3.6):

$$\log f_p(\alpha_i \mid \underline{y}) = -\log\left[(2\pi)^{(m/2)}\left|\boldsymbol{\Sigma}_i^{1/2}\right|\right] - \frac{1}{2}(\underline{y} - \underline{\mu}_i)^T|\boldsymbol{\Sigma}_i|^{-1}(\underline{y} - \underline{\mu}_i) + \log p(\alpha_i) - \log f_t(\underline{y}). \quad (3.6)$$

Koska edellisessä yhtälössä viimeinen termi f_t ($\underline{y}$) ei riipu populaatiosta, ei yhtälön viimeisenä terminä oleva logaritmoitu kokonaistodennäköisyys vaikuta vertailtavien posteriorien suuruusjärjestykseen. Täten neliöllisessä luokittelussa tutkitaan seuraavaa lauseketta (3.7), kun populaation odotusarvot ja varianssit korvataan näiden estimaateilla, eli keskiarvovektorilla $\underline{x}_i$ ja otoskovarianssimatriisilla $\mathbf{S_i}$ otoksen perusteella:

$$N_{y,i} = -\log\left[(2\pi)^{(m/2)}\left|\boldsymbol{S}_i\right|^{1/2}\right] - \frac{1}{2}(\underline{y} - \underline{x}_i)^T \boldsymbol{S}_i^{-1}(\underline{y} - \underline{x}_i) + \log \hat{p}(\alpha_i). \quad (3.7)$$

Tässä kun $\underline{y}$ on vektori, jonka populaatio ennustetaan mallilla, niin populaatioiden α_0 ja α_1 termien $N_{y,0}$ ja $N_{y,1}$ vertailu vastaa kokonaisten posterioritodennäköisyyksien vertailua. Luokitellaan havaintovektori $\underline{y}$ siihen populaatioon, kummassa edellinen lauseke antaa suuremman arvon, mikä on yhtä pitävää sen kanssa, kumman populaation estimoitu posterioritodennäköisyys on suurempi. Populaation todennäköisyyttä $p(\alpha_i)$ estimoidaan tämän osuudella otoksessa. Jos käytetään lineaarista luokittelua, oletetaan kovarianssimatriisien olevan yhtä suuria, joten kaavan (3.7) yhtälön oikean puolen ensimmäinen termi ei vaikuta suuruusjärjestykseen. Täten aiemmasta lausekkeesta saadaan tutkittavaksi lineaarisessa luokittelussa lausekkeeksi (3.8):

$$L_{y,i} = -\frac{1}{2}(\underline{y} - \underline{x}_i)^T \boldsymbol{S}^{-1}(\underline{y} - \underline{x}_i) + \log \hat{p}(\alpha_i). \quad (3.8)$$

Lineaarisessa luokitteluanalyysissä $L_{y,i}$-termien vertailu populaatioiden välillä vastaa näiden posterioritodennäköisyyksien vertailua. Luokittelun jälkeen analyysin tulosta voidaan arvioida määrittämällä, kuinka suuri osuus havainnoista luokiteltiin edellisen perusteella oikeisiin populaatioihin. Oikein ennustettujen prosenttiosuutta voidaan vertailla siihen, onko se tilastollisesti merkittävästi suurempi kuin satunnaisuus.

Jos populaation α_i osuus otoksessa on p, niin tällöin satunnaisesti luokitellessa kyseisen populaation prioritodennäköisyys on p. Merkitään symbolilla o_i, että populaatio α_i on tarkastellun havainnon todellinen populaatio ja e_i tarkoittaa, että havainto ennustettiin kuuluvan populaatioon α_i.

Kun hyödynnetään todennäköisyyslaskennan ominaisuuksia (Lindgren 1976, s. 27, 38) oikein ennustamisen todennäköisyys on esitetty kaavassa (3.9):

$$
\begin{aligned}
P('oikein') &= P((o_i \cap e_i) \cup (\neg o_i \cap \neg e_i)) = P(o_i \cap e_i) + P(\neg o_i \cap \neg e_i) \\
&= P(o_i)\,P(e_i) + P(\neg o_i)\,P(\neg e_i) \\
&= P(o_i)\,P(e_i) + (1 - P(o_i))(1 - P(e_i)) = p^2 + (1-p)^2.
\end{aligned} \tag{3.9}
$$

Jos Y on analyysin vastemuuttuja, jolla on kaksi mahdollista arvoa, otoskoko on n ja oikein ennustettujen prosenttiosuus on q, niin Y noudattaa binomijakaumaa eli Y~Bin(n,q). Tällöin binomijakauman laskusääntöjen nojalla saadaan odotusarvoksi nq ja varianssiksi nq(1-q). (Lindgren 1976, s. 163–164.)

Olkoon muuttuja Q mallin avulla oikein ennustettujen suhteellinen osuus otoksessa. Tällöin voidaan todeta myös, että oikein ennustettujen lukumäärästä otoskoolla n saadaan oikein ennustettujen suhteellisen osuuden satunnaismuuttuja Q jakamalla se otoskoolla, joten Q = Y/n. Tällöin satunnaismuuttujan odotusarvon ja varianssin arvot saadaan parametrien ominaisuuksia (Lindgren 1976, s. 124, 138) hyödyntämällä seuraavasti (3.10):

$$
\begin{aligned}
&E(Q) = E(Y/n) = E(Y)/n = nq/n = q \\
&Var(Q) = Var(Y/n) = Var(Y)(1/n)^2 = nq(1-q)(1/n)^2 = q(1-q)/n.
\end{aligned} \tag{3.10}
$$

Keskeisen raja-arvoväittämän nojalla Q noudattaa likimäärin normaalijakaumaa suurella otoskoolla n (Lindgren 1976, s. 158). Normaalijakauman parametrit odotusarvo ja varianssi määriteltiin edellä (3.10). Kun suoritetaan useita kierroksia ristiinvalidoinnissa, saadaan useita toisistaan riippumattomia oikein ennustettujen suhteellisia osuuksia. Tässä ollaan kiinnostuneita siitä, miten suuren osuuden malli ennustaa oikein tietyllä muuttujien lukumäärällä. Olkoon $Q_{b,c}$ oikein ennustettujen suhteellinen osuus ristiinvalidoinnin c:nnellä kierroksella, kun käytettyjen muuttujien lukumäärä on b. Ristiinvalidoinnin kierrosten lukumäärä on k.

Määritellään tällöin keskiarvomuuttuja $\overline{P}_i$, joka kuvaa keskimääräisesti oikein ennustettujen suhteellista osuutta i:llä muuttujalla eli $\overline{P}_i = (Q_{i,1} + \ldots + Q_{i,k})/k$. Tällöin jokaisella i:llä $\overline{P}_i$ noudattaa likimäärin normaalijakaumaa odotusarvolla $\Sigma_{j=1}(E(Q_{i,j}))$ ja varianssilla $(1/k^2)\Sigma_{j=1}(Var(Q_{i,j}))$ (Lindgren 1976, s. 124, 138).

Nyt voidaan määritellä normaalijakauman kertymäfunktion avulla 90 % luottamusväli yksittäisen tapahtuman ryhmän oikein ennustamisen todennäköisyydelle. Kun merkitään otoksesta saatua suhteellisten osuuksien keskiarvoa symbolilla $\hat{q}$’ ja estimoitua varianssia symbolilla $\hat{v}$, niin luotta-

musvälin ylä- ja alarajalle saadaan kaava (3.11): (Lindgren 1976, s. 429.)

(3.11) $$\hat{q}' \pm 1{,}6449\sqrt{\hat{\nu}}.$$

Kaavassa mainittu likiarvo 1,6449 perustuu normaalijakauman 95 % kvantiilin arvoon. Jos näin saatu luottamusvälin alaraja on otoksessa suurempi kuin aiemmin laskettu satunnaisuuden prosenttiosuus, voidaan todeta viiden prosentin riskillä, että malli ennustaa yksittäisen tapahtuman ryhmän satunnaisuutta paremmin.

3.5 Etenevä valinta erotteluanalyysille

Tässä tutkielmassa tutkitaan useiden muuttujien vaikutuksia siihen, mihin vastaajaryhmään kyseinen vastaaja kuului. Tällöin halutaan tutkia muuttujien riippuvuuksia populaatiosta sekä vertailla näitä vaikutuksia keskenään.

Käytetään etenevää valintaa, jossa valitaan malliin yksi kerrallaan paras muuttuja jäljellä olevista. Aloitetaan kokeilemalla yksitellen jokaista muuttujaa ja valitaan näistä paras, kun käytetään arviointikriteerinä sitä, miten suuri mallin luokittelutarkkuus on. Tämän valinnan jälkeen valitaan seuraavaksi paras muuttuja samalla arviointikriteerillä, kun malliin kuuluu myös sinne jo aiemmin valitut muuttujat. (Elisseeff & Guyon 2003.)

Jatketaan muuttujien poimimista niin kauan, kunnes kaikki muuttujat ovat mallissa. Oletetaan edelleen, että vastemuuttuja Y saa joko arvon 0 tai 1, ja merkitään vastemuuttujan mukaisesti jaettuja populaatioita vastaavasti α_0 ja α_1. Olkoon $\underline{x}_i$ on yksi mallissa käytetty havaintovektori ja $p(\alpha_1 \mid \underline{x}_i)$ posterioritodennäköisyys sille, että kyseinen havainto kuuluu populaatioon α_1. Termin y_i arvo on 1, mikäli havainto kuuluu populaatioon α_1, muuten se on nolla.

Lasketaan nyt mallin luokittelutarkkuus siten, että jos havainto kuuluu analyysin perusteella todelliseen populaatioon suurella estimoidulla posterioritodennäköisyydellä, kasvattaa tämä paljon tarkkuutta. Vastaavasti jos havainto kuuluu mallin perusteella vain pienellä posterioritodennäköisyydellä todelliseen populaatioonsa, vähentää tämä paljon tarkkuutta. Täten ”pehmeäksi luokittelutarkkuudeksi” saadaan kaava (3.12):

(3.12) $$\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot \hat{p}(\alpha_1 | \underline{x}_i) + (1-y_i)\cdot(1-\hat{p}(\alpha_1 | \underline{x}_i))\right].$$

Se muuttuja, jonka lisääminen malliin antaa korkeimman luokittelutarkkuuden opetusdatassa, valitaan seuraavaksi. Näin malliin lisätään lopulta kaikki muuttujat. Mallien tehokkuutta voidaan arvioida luokittelemalla havaintoja testidatassa, jota ei käytetty mallien valinnassa.

Koska testidatassa luokittelu riippuu datan valinnasta, liittyy tähän jonkin verran satunnaisuutta. Käytetään tämän vuoksi ristiinvalidointia etenevälle valinnalle ja luokittelulle. Tällöin data jaetaan t:hen osadataan D_1, D_2, … D_t. Ensimmäisessä vaiheessa suoritetaan etenevä valinta opetusdatassa, joka on alkuperäinen data, josta on poistettu D_1. Tämän jälkeen suoritetaan luokittelu testidatassa, joka on ainoastaan osadata D_1. Vastaavasti suoritetaan etenevä valinta ja luokittelu jokaisella osadatalla vaihtamalla käytettyä opetus- ja testidataa. Näin määritellään etenevän valinnan mallit sekä mallien oikein ennustamien ryhmien prosenttiosuudet, jotka ovat toisistaan riippumattomia osadatojen ollessa erillisiä.

Luokitteluanalyysissä, jossa ei tutkita oikein ennustettujen prosenttiosuuksia, vaan keskitytään muuttujien tärkeysjärjestykseen, lisätään todellisten muuttujien lisäksi generoitu satunnaismuuttuja. Tämän avulla voidaan arvioida sitä, miten paljon paremmin kyselyn kysymyksiin perustuvat muuttujat selittivät vastaajan ryhmää verrattuna satunnaisuuteen. Lisäksi tällöin käytetään ristiinvalidoinnin yksinkertaisempaa muotoa, jossa ensin suoritetaan etenevä valinta datassa D_1, sitten datassa D_2 ja niin edelleen käyden läpi kaikki erilliset osadatat. Tällöin kaikki mallin opetuksessa käytetyt osadatat ovat toisistaan riippumattomia.

3.6 Merkkitesti

Tutkitaan vähintään ordinaalisia muuttujia Y_1 ja Y_2. Muuttujat ovat pareittaisia, tässä ne ovat aineiston muuttujan ja generoidun satunnaismuuttujan sijoitukset samassa ristiinvalidoinnin kierroksessa, kun Y_2 on generoitu muuttuja. Oletetaan muuttujien olevan periaatteessa jatkuvia, vaikka näitä mitattaisiinkin yksinkertaisemmin diskreeteillä arvoilla. Tavoitteena olisi tutkia sitä, sijoittuuko aineiston muuttuja keskimääräisesti paremmin kuin generoitu satunnaismuuttuja. Käytetyssä ristiinvalidoinnissa suoritetaan n kierrosta. Määritellään apumuuttuja L, joka saa arvon 1 kun $Y_1<Y_2$ ja muuten arvon 0. Tällöin jos arvon 1 todennäköisyys on p_K ja L:n arvo määritetään n kertaa toisistaan riippumattomissa satunnaiskokeissa, niin arvojen summamuuttuja K noudattaa binomijakaumaa parametreilla p_K ja n. (Lindren, 1976 s. 163, 505–507.)

Jos nyt aineiston muuttuja Y_1 sijoittuu generoitua muuttujaa paremmin ristiinvalidoinnin tutkittavalla kierroksella, on $Y_1<Y_2$ ja myös L=1. Jos aineiston muuttuja sijoittuu odotusarvoisesti generoitua muuttujaa paremmin, on tällöin $p_K>0,5$, koska tällöin tapahtuma L=1 on todennäköisempi

kuin sen komplementti. Määritellään tällöin hypoteesit:

H_0: $p_K = 0{,}5$
H_1: $p_K > 0{,}5$

Määritellään nyt merkkitesti hypoteesien tarkasteluun. Koska nollahypoteesin ollessa tosi K noudattaa binomijakaumaa parametreilla n ja 0,5, voidaan määrittää testauksen p-arvo havaitulle K:n arvolle k binomijakauman todennäköisyysfunktion avulla kuten kaavassa (3.13):

(3.13) $$p = \sum_{i=k}^{n} \binom{n}{i} 0{,}5^i (1-0{,}5)^{n-i} = \sum_{i=k}^{n} \binom{n}{i} 0{,}5^n$$

Nyt merkkitestissä ei ole tarpeen tehdä muita oletuksia muuttujien käyttäytymisestä. (Lindren, 1976 s. 163, 505–507.)

Tässä tutkielmassa etenevän valinnan tuloksena saadaan muun muassa muuttujien sijoitukset siitä, millä askeleella nämä valittiin malliin. Tarkoituksena on vertailla, valitaanko nämä keskimääräisesti aikaisemmin kuin generoitu satunnaismuuttuja. Kun tutkitaan yksittäistä aineiston muuttujaa, saadaan jokaisella ristiinvalidoinnin kierroksella pari, jossa on aineiston muuttujan sekä generoidun muuttujan sijoitus. Merkkitestiä varten tehtävässä etenevässä valinnassa ristiinvalidoinnin opetusosadatat ovat erillisiä, joten näin saatavat parit muuttujien sijoituksista eri ristiinvalidoinnin kierroksilla ovat toisistaan riippumattomia. Näiden parien analysoinnissa voidaan käyttää merkkitestiä sen tutkimiseen, onko sijoituksissa tilastollisesti merkitsevää eroa havaittavissa.

3.7 Pääkomponenttianalyysi ja Sammon kartta

Oletetaan että tutkittavana on d-ulotteinen avaruus, joka koostuu d:stä muuttujasta X_1, …, X_d. Oletetaan lisäksi, että muuttujat on skaalattu siten, että jokaisen muuttujan keskiarvo on nolla ja varianssi 1. Pääkomponenttianalyysissä on tavoitteena löytää muuttujille sellainen lineaarikombinaatio, että sillä pystytään selittämään aineistossa havaittua vaihtelua mahdollisimman hyvin. Tällöin saadaan tietoa tiivistettyä vähentämällä aineiston muuttujien lukumäärää käyttämällä ainoastaan k:ta ensimmäistä pääkomponenttia, kun k < d. (Johnson & Wichern 2007.)

Olkoon tuntematon d x d -kerroinmatriisi $\mathbf{A}=[a_{ij}]$. Tällöin jos tutkitaan aineistoon kuuluvaa ha-

vaintovektoria $X^T = [x_1, \ldots, x_n]$, ovat tätä vastaavat lineaarikombinaatiot $Y_1, Y_2, \ldots, Y_d$ muotoa:

$Y_i = \underline{a}_i X = a_{i1}X_1 + a_{i2}X_2 + \ldots a_{ip}X_n \quad$ kun i = 1, 2,…, d

Tällöin lineaarikombinaatioita Y_i kutsutaan pääkomponenteiksi. Kun tavoitteena on sellaiset lineaarikombinaatiot, että ne selittäisivät mahdollisimman hyvin alkuperäisen aineiston vaihtelua, määritellään matriisin **A** kertoimet siten, että:

1) $\sum_{i=1}^{p} a_{ji}^2 = 1$ kaikilla j=1, 2,..., p

2) $\mathrm{Var}(Y_1)$ on mahdollisimman suuri

3) $\mathrm{Var}(Y_i)$ on mahdollisimman suuri, kun i=2,…, p ja $\mathrm{Cov}(Y_i, Y_j) = 0$ jokaisella j:llä $1 \leq j < i$

Kun $\boldsymbol{\Sigma} = \mathrm{Cov}(X)$, tehdään kovarianssimatriisille ominaisarvohajotelma $\boldsymbol{\Sigma} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^T$, jossa **V** kuvaa ominaisvektorimatriisia ja $\boldsymbol{\Lambda}$ on diagonaalimatriisi ominaisarvoilla. Nyt kerroinmatriisi **A** saadaan hajotelmasta, sillä $\mathbf{A} = \mathbf{V}^T$. Tällöin voidaan laskea aiemman kaavan avulla pääkomponentit, joiden avulla voidaan kuvata alkuperäistä aineistoa vähemmällä määrällä muuttujia. (Johnson & Wichern 2007.)

Käytetään lisäksi tiedon tiivistämiseen Sammon karttaa. Tutkittavana on edelleen aiemmin esitelty aineisto. Kun aineistoon kuuluu n kappaletta d-ulotteisia havaintoa $X'_1, \ldots, X'_n$, niin tavoitteena on löytää sellainen tiivistetty aineisto **Y'**, että alkuperäisen aineiston etäisyydet säilyisivät vastaavina. Määritellään **Y'**:n vektorit $Y'_1, \ldots, Y'_n$ kaksiulotteisiksi. Merkitään alkuperäisen aineiston etäisyyksiä $d_{ij} = \mathrm{dist}(X'_i, X'_j)$ ja tiivistetyn aineiston etäisyyksiä $d_{ij}^{\circ} = \mathrm{dist}(Y'_i, Y'_j)$. Tällöin määritellään virhe E kuten kaavassa (3.14):

(3.14)
$$E = \frac{1}{\sum_{i<j} d_{ij}^{\circ}} \sum_{i<j}^{n} \frac{(d_{ij}^{\circ} - d_{ij})^2}{d_{ij}^{\circ}}.$$

Tavoitteena on tällöin minimoida virhe, jolloin etäisyydet tiivistetyssä datassa ovat vastaavia kuin alkuperäisessä datassa. (Sammon 1969.)

Käytetään ensimmäisenä tiivistettynä **Y'**-datana pääkomponenttianalyysin kahta ensimmäistä pääkomponenttia. Tavoitteena olisi löytää iteroimalla aina edellistä **Y'**-dataa parempi uusi tiivistetty data, kun hyvyyden mittarina käytetään kaavan (3.14) antamaa virhettä.

Tällöin käytetään Sammon kartassa virheen minimoimiseen NLM-algoritmia (nonlinear map-

ping), jossa siirrytään negatiivisen gradientin suuntaan tiivistetyssä datassa. Tällöin pyritään löytämään virheen paikallinen minimi. Näin jatketaan niin kauan, kunnes karttaa ei pystytä enää merkittävästi parantamaan, ja näin saadaan tuloksena tiivistetty aineisto. Toisin kuin pääkomponenttianalyysissä, Sammon kartassa ei vaadita millään askeleella **Y'**:n vektoreiden olevan lineaarikombinaatioita alkuperäisen aineiston vektoreista, vaan myös epälineaarisen muunnokset ovat mahdollisia. (Sammon 1969.)

4 Aineiston analysointi

Seuraavaksi analysoidaan tutkimusaineistoa. Analyysi aloitetaan tutkimalla kaikkia kolmea vastaajaryhmää yhtäaikaisesti. Tässä käytetään Kruskal-Wallisin testiä, kun testataan muuttujien jakaumien yhtäsuuruutta ryhmien välillä. Lisäksi testin aputuloksia tarkastellaan graafisesti.

Tämän jälkeen alaluvussa 4.2 tutkitaan ryhmien välillä havaittavia eroja, kun vertaillaan lopettaneiden ryhmiä erikseen kaikkiin tutkimuksiin vastanneiden kanssa. Menetelmänä tässä käytetään erottelu- ja luokitteluanalyysiä sekä näiden tulosten tulkinnassa myös merkkitestiä. Tällöin menetelmissä käytetään hyväksi tietoja vastaajien todellisista vastaajaryhmistä. Lisäksi käytetään Sammon karttaa erilaisten vastaajien ryhmien erotteluun, kun tässä ei käytetä hyväksi tietoa siitä, lopettiko vastaaja vastaamisen kesken.

Alaluvun 4.2 analyyseihin käytettiin R-ohjelmistoa, ja alaluvussa 4.1 sekä R- että SPSS-ohjelmistoa.

4.1 Kaikki vastaajaryhmät erikseen

Tutkitaan seuraavassa alaluvussa kolmea vastaajaryhmää yhtäaikaisesti. Tällöin vertaillaan sitä, ovatko eri ryhmien vastanneet ensimmäisessä kyselyssä samoin.

4.1.1 Muuttujien jakaumien yhtäsuuruus vastaajaryhmien välillä

Seuraavaksi tutkitaan, onko vastauksissa havaittavissa eroja eri vaiheissa lopettaneiden ja niiden välillä, jotka vastasivat kaikkiin kyselyihin. Analyysissä käytetään Kruskal-Wallisin testiä, joka on epäparametrinen yleistys yksisuuntaisesta varianssianalyysistä. Tässä nollahypoteesina on, että tut-

kittavan muuttujan jakaumat ovat samoja kolmessa eri ryhmissä. Vaihtoehtoisena hypoteesina on, että ryhmien välillä on havaittavissa eroja.

Tarkastelussa käytetään lomakkeen kysymyksiä, jotka esiteltiin aineiston kuvailussa, mutta otetaan ainoastaan vastaajien ensimmäiset vastaukset, sillä ainoastaan näihin kaikkien ryhmien vastaajat ovat voineet vastata. Lisäksi tässä käytetään vastaajien alkuperäisiä vastauksia, joita ei ole luokiteltu. Näitä käytetään, sillä Kruskal-Wallisin testi vaatii tarkasteltavaksi vähintään järjestysasteikollisen muuttujan, mutta tämän ei ole välttämätöntä olla numeerinen. Tällöin järjestysasteikolliset muuttujat sekä diskreettejä lukuja saavat muuttujat oletetaan luonteeltaan jatkuviksi, vaikka näitä mitataan ainoastaan diskreeteillä arvoilla.

Suoritetaan testaus erikseen jokaiselle selittävälle muuttujalle. Saadaan testisuureen arvot ja p-arvot, jotka on koottu taulukkoon 4.1. 5 % riskitasolla merkittävät testisuureen arvot on korostettu keltaisella.

Taulukko 4.1. Kruskal-Wallisin testien testisuureiden arvot sekä p-arvot.

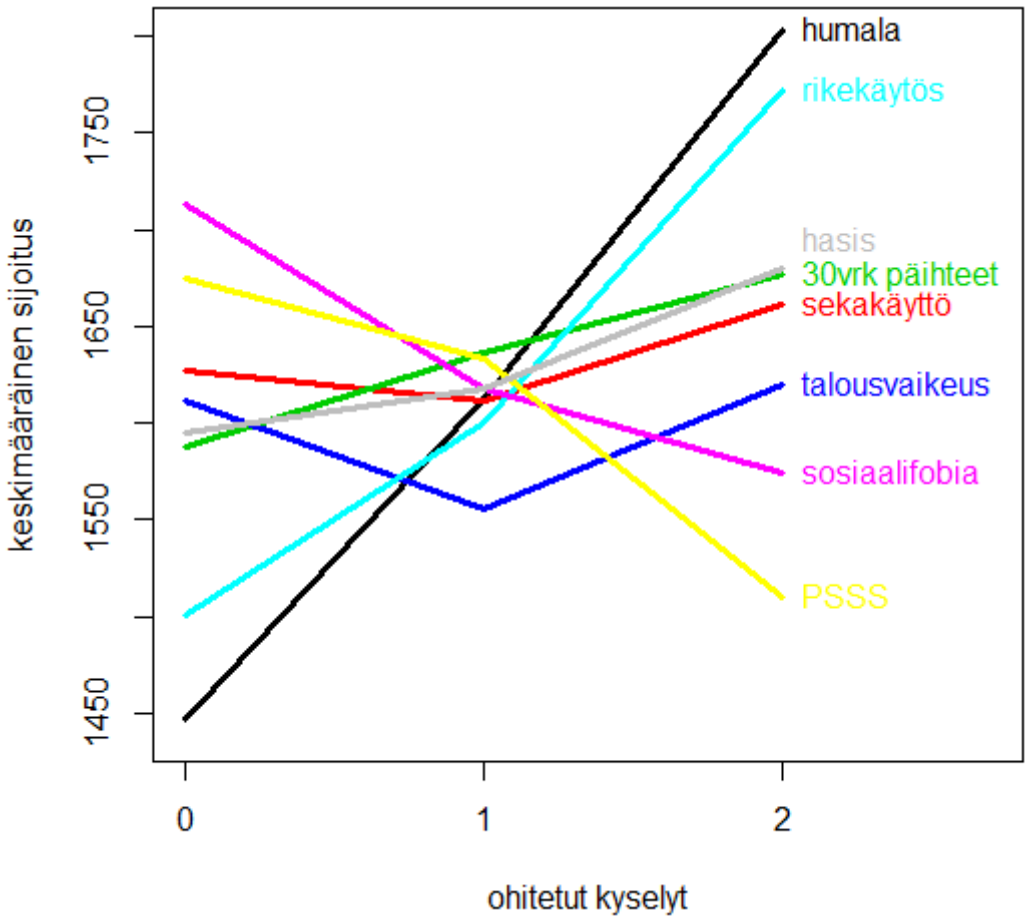
	χ^2-testisuureen arvo	p-arvo
Koettu terveys	1,636	0,441
Humalahakuinen juominen	92,563	< 0,001
Päihteiden sekakäyttö	6,040	0,049
Kovien huumeiden käyttö	1,817	0,403
Huumeiden käyttö viimeisenä kuukautena	10,458	0,005
Taloudelliset vaikeudet	6,469	0,039
Masennus	1,203	0,548
Rikekäyttäytyminen	48,720	< 0,001
Aggressiivinen käyttäytyminen	1,109	0,574
Sosiaalinen ahdistuneisuushäiriö	12,487	0,002
Sosiaalinen tuki	19,134	< 0,001
Hasiksen käyttö	14,450	0,001

Taulukosta nähdään, että kaikkien päihteiden käyttöön liittyvien muuttujien, lukuun ottamatta kovia huumeita, p-arvot ovat pienempiä kuin 0,05. Täten näissä nollahypoteesi voidaan hylätä 5 % riskillä ja todeta, että muuttujien jakaumat eroavat toisistaan ryhmien välillä. Vastaava päätelmä voidaan tehdä seuraaville muuttujille: taloudelliset vaikeudet, rikekäyttäytyminen, sosiaalinen ahdistuneisuushäiriö ja sosiaalinen tuki.

Sen sijaan koetun terveyden, kovien huumeiden käytön, masennuksen ja aggressiivisen käytök-

sen nollahypoteesit hyväksytään 5 % riskillä, joten näiden voidaan olettaa jakautuvan samoin kaikissa ryhmissä.

Havainnollistetaan ryhmien eroavaisuuksia ehdollisilla tunnusluvuilla. Aiemmin todettiin humalahakuisen juomisen poikkeavan ryhmien välillä. Käytetään aiemmin laskettuja tuloksia, kun määriteltiin vastausten suuruusjärjestys testaamista varten. Tunnusluku ”keskimääräinen sijoitus” perustuu testauksen aputuloksina käytettyihin tietoihin havaintojen suuruusjärjestyksestä. Aineistosta voidaan laskea ryhmille *kaikkiin vastanneet*, *vain kolmannen yli hypänneet*, ja *toisen yli hypänneet* tunnusluvuiksi 1447,43, 1612,52 ja 1802,93. Tässä muuttujan pieni arvo tarkoittaa humalahakuisen juomisen olevan vähäistä. Voidaan havaita, että humalahakuisessa juomisessa näyttäisi olevan trendiä, joka kasvaa yli hypättyjen kyselyjen kasvaessa. Havainnollistetaan otoksesta laskettuja keskimääräisiä suuruusjärjestyksen sijoituksia kuviolla 4.1, kun tutkitaan niitä muuttujia, joissa aiemmin havaittiin eroja ryhmien välillä.



Kuvio 4.1. Kruskal-Wallisin testissä käytetyt keskimääräistä sijoitusta suuruusjärjestyksessä kuvaavat tunnusluvut eri muuttujilla eri vastaajaryhmien välillä. Kuviossa toisen kyselyn ohittaneet ovat oikeanpuolimmaisina, vaikka jotkut näistä vastasivat viimeiseen kyselyyn.

Kuvion perusteella otoksessa näyttäisi olevan nousevaa trendiä humalahakuisen juomisen lisäksi myös muuttujissa rikekäyttäytyminen, hasiksen käyttö ja päihteiden käyttö kuukauden aikana. Laskevaa trendiä näyttäisi olevan taas muuttujissa PSSS ja sosiaalifobia. Kyselyssä huomattavaa on, että muuttujilla korkeita arvoja voidaan pitää ”kielteisinä” lukuun ottamatta muuttujaa PSSS, jossa suuret arvot kuvaavat hyvää sosiaalista tukea. Tällöin myös kuviossa ilmoitettuja keskimääräistä sijoitusta kuvaavia arvoja tulkitaan samalla tavalla, joten esimerkiksi humalahakuinen juominen vaikuttaisi lisääntyvän, kun ohitettuja kyselyjä on useampia.

4.2 Lopettaneiden ryhmät erikseen vs. kaikkiin vastanneet

Tässä alaluvussa tutkitaan haastateltavien vastausten riippuvuutta vastaajaryhmästä, kun keskeyttäneiden ryhmien vertaillaan erikseen kaikkiin kyselyihin vastanneiden kanssa. Lisäksi kyselyiden vastauksia havainnollistetaan Sammon kartalla kahdessa ensimmäisessä kyselyssä erikseen.

4.2.1 Erottelu- ja luokitteluanalyysi, ensimmäisessä vaiheessa lopettaneet vs. kaikkiin vastanneet

Jatketaan aineiston analysointia tarkastelemalla muuttujien yhteyttä vastaajaryhmän kanssa. Tässä käytetään erottelu- ja luokitteluanalyysiä, kun pääpaino on luokitteluanalyysissä. Verrataan ensin ensimmäisessä vaiheessa lopettaneita niihin, jotka vastasivat kaikkiin kyselyihin. Nyt käytetään luokitteluanalyysiä etenevällä valinnalla, kun valitaan malliin uusi muuttuja pehmeän luokittelutarkkuuden perusteella. Lisätään näin kaikki käytetyn aineiston muuttujat lopulta malliin.

Käytetään varsinaisen datan lisäksi generoitua satunnaismuuttujaa, jonka avulla arvioidaan valittujen muuttujien tärkeyttä mallissa. Kun käytetään ristiinvalidointia, saadaan eri osadatoille eri muuttujavalinnat. Määritellään jokaiselle muuttujalle keskimääräinen sijoitus sekä sen hajonta yli ristiinvalidoinnin kierrosten. Tehdään nämä tarkastelut sekä lineaarisella että neliöllisellä luokitteluanalyysillä. Tässä vaiheessa käytetään yksinkertaisempaa ristiinvalidointia, jossa jokainen opetusosadata on toisistaan erillinen. Tulokset on koottu taulukoihin 4.2 ja 4.3.

Luokitteluanalyysien tulosten perusteella testataan miten hyvin valitut muuttujat ennustavat haastatellun vastaajaryhmää. Käytetään tässä pareittaisille havainnoille merkkitestiä, kun verrataan yksi kerrallaan jokaista tutkittua muuttujaa generoituun muuttujaan. Tämä tarkastelu tehdään erikseen molempien analyysien tuloksille. Tällöin nollahypoteesina on, että tutkittava muuttuja ja generoitu muuttuja sijoittuvat odotusarvoisesti yhtä hyvin. Vaihtoehtoisena hypoteesina on, että generoi-

dun muuttujan keskimääräinen sijoitus on parempi kuin tutkittavan muuttujan. Testausten näin saadut p-arvot on myös listattu taulukoihin 4.2 ja 4.3. Molemmissa taulukoissa tulokset on järjestetty keskimääräisen sijoituksen mukaan, joka on listattu sarakkeessa ”ka”. Sijoituksen keskihajonta on sarakkeessa ”kh” ja testauksesta saatu p-arvo on sarakkeessa ”p”. P-arvot on ilmoitettu kolmen desimaalin tarkkuudella.

Taulukot 4.2. ja 4.3. Lineaarisen luokitteluanalyysin tulokset ovat vasemmanpuoleisessa taulukossa ja neliöllisen luokitteluanalyysin oikeassa. Jokaiselle muuttujalle on laskettu sijoituksen keskiarvo ja keskihajonta yli ristiinvalidoinnin kierrosten. Lisäksi on testattu, eroaako muuttujien sijoitus generoidun muuttujan sijoituksesta. Merkkitestauksen p-arvot on lueteltu sarakkeissa ”p”.

muuttuja	ka	kh	p
humalajuominen	1.89	1.45	0.002
sosiaalifobia	5.44	3.00	0.020
PSSS	5.44	4.00	0.090
päihteet 30vrk	6.33	3.32	0.020
rikekäytös	6.89	4.17	0.090
aggressiivisuus	7.11	3.18	0.090
voimakkaat huumeet	7.56	3.36	0.254
masennus	7.56	4.69	0.254
mielipide terveydestä	7.67	2.40	0.020
talousongelmat	8.00	3.81	0.090
sekakäyttö	8.11	3.44	0.254
hasis	8.56	3.54	0.254
generoitu	10.44	2.46	1.000

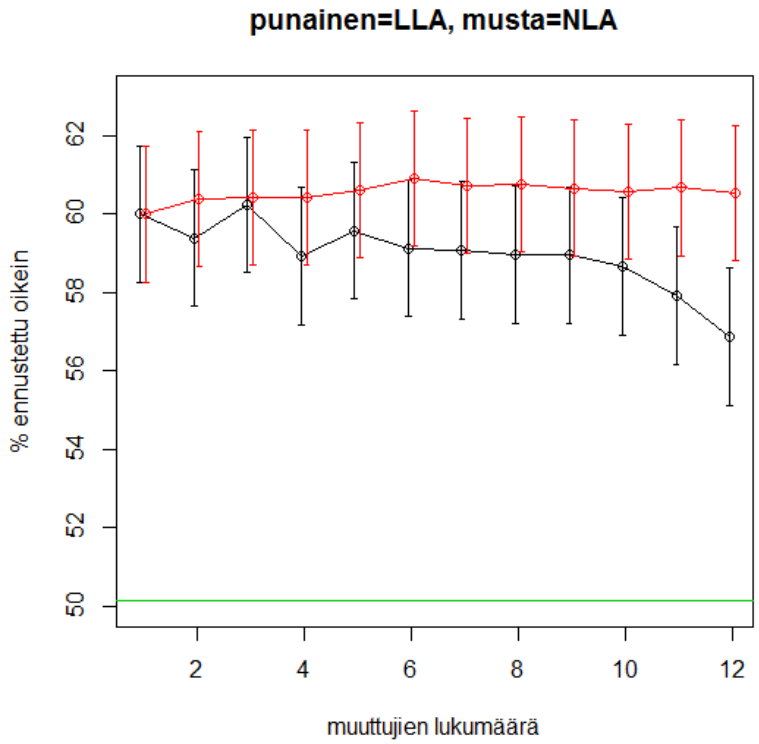
muuttuja	ka	kh	p
humalajuominen	3.56	3.21	0.020
PSSS	4.44	3.21	0.020
aggressiivisuus	4.78	1.86	0.090
masennus	6.00	2.74	0.090
rikekäytös	6.22	4.24	0.254
sosiaalifobia	6.33	3.32	0.500
päihteet 30vrk	6.44	3.94	0.020
hasis	6.67	3.28	0.020
mielipide terveydestä	8.22	3.07	0.500
talousongelmat	8.56	2.55	0.500
generoitu	9.00	2.83	1.000
sekakäyttö	9.22	3.77	0.746
voimakkaat huumeet	11.56	3.61	0.998

Taulukoista voidaan havaita, että molemmissa luokitteluanalyyseissä satunnaisuutta tilastollisesti merkittävästi paremmin sijoittuneita muuttujia ovat humalahakuinen juominen, PSSS, aggressiivinen käyttäytyminen ja huumausaineiden käyttö viimeisen kuukauden aikana, kun testauksen riskitasona on 10 %. Lisäksi muuttujat sosiaalifobia, rikekäyttäytyminen, mielipide omasta terveydentilasta, taloudelliset ongelmat, masennus ja hasiksen käyttö todettiin merkitseviksi toisessa tarkasteluista. Täten siis ainoastaan sekakäyttö ja voimakkaat huumausaineet olivat ei-merkitseviä molemmissa tarkasteluissa tällä riskitasolla.

Tarkastellaan lisäksi mallien oikein ennustettujen prosenttiosuuksia. Tämä tehdään ristiinvalidoinnin jokaisella kierroksella erikseen sen hetkiselle testidatalle. Tässä käytetään ”tavallista” ristiinvalidointia, jossa valitaan yksi kerrallaan jokainen osadata testidataksi, jolloin muut havainnot muodostavat opetusdatan. Testidatat ovat siis toisistaan erillisiä. Käytetään tässä ainoastaan varsinaista dataa ilman generoitua satunnaismuuttujaa. Tällöin ei suoriteta merkkitestausta, vaan keskitytään testidatan ryhmien ennustamiseen.

Yksittäisen havainnon oikein ennustamisen todennäköisyydelle voidaan laskea normaalijakau-

ma-approksimaatiolla 90 % luottamusväli, jonka alarajasta ollaan kiinnostuneita. Lasketaan keskimääräiselle oikein ennustettujen prosenttiosuudelle kyseinen luottamusväli. Havainnollistetaan näitä tuloksia kuviolla 4.2. Kuviossa pallot kuvaavat keskimääräistä oikein ennustettujen prosenttiosuutta otoksessa kyseisellä muuttujien lukumäärällä, ja janat kuvaavat 90 % luottamusväliä. Vihreällä on kuvattu satunnaisesti oikein ennustettujen teoreettinen prosenttiosuus, joka oli noin 50,13. Lineaarinen luokitteluanalyysi on merkitty punaisella ja neliöllinen mustalla.



Kuvio 4.2. Luokitteluanalyysien mallien oikein ennustettujen ryhmien prosenttiosuuksia sekä näiden 90 % luottamusvälejä. Satunnaisesti oikein ennustettujen suhteellinen osuus on merkitty vihreällä viivalla.

Kuvion perusteella huomataan sekä lineaarisen että neliöllisen analyysin mallien ennustavan yksittäisen havainnon populaatio aina satunnaisuutta paremmin 5 % riskillä. Oikein ennustettujen prosenttiosuudet olivat otoksessa hieman korkeammat lineaarisella luokittimella. Jos verrataan neliöllisen luokittelun otoksesta laskettua luottamusväliä lineaarisen luokittelun luottamusväliin, voidaan todeta luottamusvälien menevän aina päällekkäin lukuun ottamatta 12 muuttujan mallia. Näyttäisi siis, että kumpikaan luokittimista ei pääasiassa anna toista merkittävästi parempia ennusteita. Lisäksi neliöllisen luokittelun tulos otoksessa näyttäisi heikkenevän mallin kasvaessa, kun taas lineaarinen luokittelu parantuu hieman ja antaa sitten kutakuinkin samaa tulosta.

4.2.2 Erottelu- ja luokitteluanalyysi, toisessa vaiheessa lopettaneet vs. kaikkiin vastanneet

Tehdään seuraavaksi vastaavat tarkastelut kuin alaluvussa 4.2.1, kun vertaillaan toisessa vaiheessa lopettaneita kaikkiin vastanneisiin. Täten käytössä on enemmän tietoa kuin edellä, sillä tutkittavassa osajoukossa haastateltavat vastasivat sekä ensimmäisen että toisen kyselyn kysymyksiin. Nyt luokitteluanalyyseistä saadaan seuraavat taulukot 4.4 ja 4.5. Taulukoiden tulokset on järjestetty keskimääräisen sijoituksen mukaan, joka on listattu sarakkeessa ”ka”. Sijoituksen otoskeskihajonta on sarakkeessa ”kh” ja testauksesta saatu p-arvo on sarakkeessa ”p”. Saatu p-arvo on ilmoitettu kolmen desimaalin tarkkuudella. Varsinaisia tutkimuskysymyksiä kuvaavien muuttujien ohella analyysissä käytettiin generoitua satunnaismuuttujaa, jotta varsinaisten muuttujien tärkeyttä mallissa voidaan arvioida. Tämä muuttuja on nimetty ”generoitu”.

Taulukot 4.4. ja 4.5. Lineaarisen luokitteluanalyysin tulokset ovat vasemmanpuoleisessa taulukossa ja neliöllisen luokitteluanalyysin oikeanpuolimmaisessa. Jokaiselle muuttujalle on laskettu sijoituksen keskiarvo ja keskihajonta yli ristiinvalidoinnin kierrosten. Lisäksi on testattu, eroaako muuttujien sijoitus generoidun muuttujan sijoituksesta, ja testauksen p-arvot on listattu. Muuttujan nimessä järjestysluku 1. viittaa ensimmäiseen kyselyyn ja 2. toiseen.

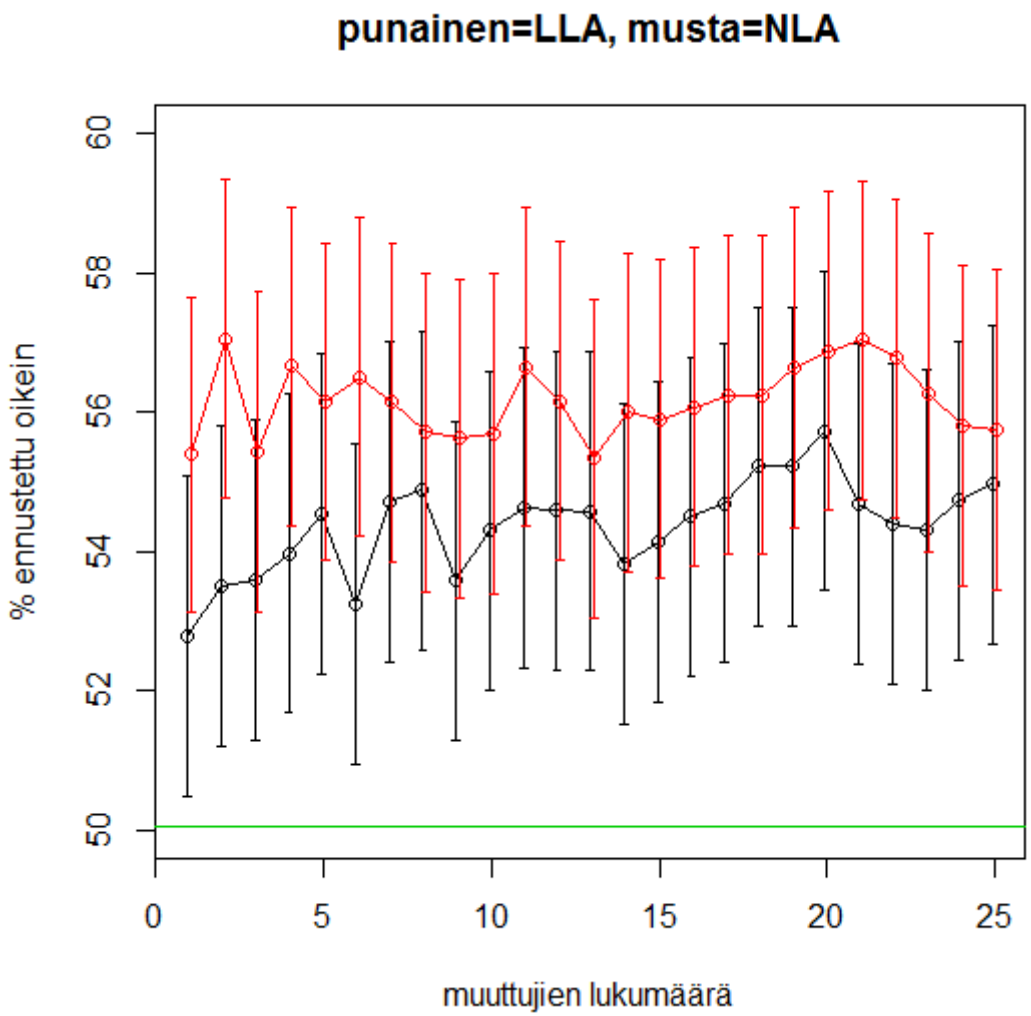
muuttuja	ka	kh	p
2. aggressiivisuus	7.78	5.36	0.002
toimettomuus	8.44	5.48	0.002
1. talousongelmat	8.89	7.04	0.090
1. humalajuominen	9.11	7.62	0.002
1. masennus	10.67	6.44	0.090
1. sekakäyttö	11.44	7.99	0.090
1. sosiaalifobia	11.89	6.62	0.090
2. humalajuominen	12.11	10.30	0.090
2. sosiaalifobia	12.22	7.33	0.500
2. rikekäytös	12.44	8.23	0.254
2. voimakkaat huumeet	12.78	6.38	0.090
2. hasis	12.89	6.55	0.090
2. sekakäyttö	13.33	7.87	0.500
2. PSSS	13.67	6.26	0.090
1. mielipide terveydestä	14.22	7.73	0.500
1. aggressiivisuus	14.44	7.58	0.500
2. päihteet 30vrk	14.44	8.35	0.254
1. hasis	14.78	7.68	0.254
1. PSSS	15.11	7.69	0.254
2. masennus	15.44	5.77	0.500
1. päihteet 30vrk	16.56	7.73	0.500
1. voimakkaat huumeet	16.67	7.91	0.500
2. mielipide terveydestä	16.89	9.05	0.910
generoitu	18.11	4.04	1.000
1. rikekäytös	18.22	6.10	0.500
2. talousongelmat	18.44	8.02	0.746

muuttuja	ka	kh	p
2. humalajuominen	7.00	5.62	0.016
2. päihteet 30vrk	7.17	3.06	0.016
1. humalajuominen	7.17	4.54	0.109
1. hasis	8.00	5.29	0.344
2. rikekäytös	9.33	6.35	0.109
1. PSSS	9.83	5.38	0.109
2. aggressiivisuus	10.00	5.93	0.656
1. masennus	10.33	7.06	0.344
1. päihteet 30vrk	12.00	5.83	0.344
2. PSSS	12.00	6.48	0.656
2. mielipide terveydestä	13.33	6.25	0.344
1. sekakäyttö	13.50	8.83	0.656
1. aggressiivisuus	13.67	6.28	0.109
1. rikekäytös	14.00	7.27	0.344
1. talousongelmat	14.17	6.11	0.656
2. sosiaalifobia	14.33	7.06	0.891
generoitu	14.50	3.45	1.000
1. sosiaalifobia	14.67	8.31	0.891
1. mielipide terveydestä	14.67	9.22	0.656
2. hasis	14.83	7.94	0.656
2. masennus	14.83	9.70	0.891
toimettomuus	16.00	11.08	0.891
2. talousongelmat	16.50	7.40	0.891
2. sekakäyttö	20.50	3.15	0.984
1. voimakkaat huumeet	24.17	2.14	1.000
2. voimakkaat huumeet	24.50	1.38	1.000

Tuloksista havaitaan niiden eroavan jokin verran aiemman alaluvun tuloksista. Tällä kertaa molemmissa analyyseissä satunnaisuutta paremmin sijoittui ainoastaan toisen kyselyn humalahakuinen juominen, kun käytetään 10 % riskitasoa. Ainoastaan toisessa analyysissä keskimääräisesti satunnaisuutta paremmin sijoittuivat toisen kyselyn aggressiivinen käytös, toimettomuus, voimakkaat huumeet, hasiksen käyttö, sosiaalinen tuki ja viimeisen kuukauden päihteiden käyttö. Lisäksi ensimmäisen kyselyn talousongelmat, humalahakuinen juominen, masennus, päihteiden sekakäyttö ja sosiaalifobia olivat ainoastaan toisessa analyysissä merkitseviä, kun riskitasona käytetään edelleen 10 %.

Tuloksista voidaan myös huomata, että neliöllisessä luokitteluanalyysissä molemmat tilastollisesti merkittävistä muuttujista olivat toisesta kyselystä. Tämän perusteella vaikuttaisi, että tässä analyysissä tuoreemmat vastaukset ennustivat paremmin tulevaa vastauskäyttäytymistä. Sen sijaan lineaarisessa luokitteluanalyysissä tilastollisesti merkitsevistä muuttujista viisi oli ensimmäisestä kyselystä ja kuusi oli jälkimmäisestä. Lisäksi vaikuttaisi, että generoitu satunnaismuuttuja sijoittuu heikommin lineaarisessa kuin neliöllisessä luokitteluanalyysissään jäädessään kolmanneksi viimeiseksi.

Tutkitaan vielä testidatassa miten luokitteluanalyysin mallit ennustavat testidatassa haastatellun vastaajaryhmän. Käytetään tässä tavallista ristiinvalidointia osittain päällekkäin menevillä opetusdatoilla, lisäksi analyysissä käytetään ainoastaan varsinaista dataa ilman generoitua muuttujaa. Otoksessa oikein ennustettujen prosenttiosuuksista sekä niiden 90 % luottamusväleistä saadaan kuvio 4.3. Kuviossa pallot kuvaavat keskimääräistä oikein ennustettujen prosenttiosuutta otoksessa kyseisellä muuttujien lukumäärällä. Lineaarinen luokitteluanalyysi on väritetty punaisella ja neliöllinen mustalla. Vihreä viiva kuvaa satunnaisesti oikein ennustettujen prosenttiosuutta, joka oli tässä noin 50,05.



Kuvio 4.3. Luokitteluanalyysien mallien oikein ennustettujen ryhmien keskimääräiset prosenttiosuudet sekä näiden 90 % luottamusvälit, kun ryhmitellään mallit ensin niiden muuttujien lukumäärän perusteella. Satunnaisesti oikein ennustettujen osuus on kuvattu vihreällä viivalla.

Kuviosta 4.3 havaitaan molempien mallien antavan keskimäärin satunnaisuutta parempia ennusteita jokaisella muuttujien lukumäärällä, kun käytetään 5 % riskitasoa. Tässä vertaillaan satunnaisesti oikein ennustettujen osuutta luottamusvälin alarajaan. Otoksessa lineaarinen luokitteluanalyysi antoi hieman parempia tuloksia, mutta luottamusvälejä tutkimalla havaitaan, että erot neliöllisen luokittelun prosenttiosuuteen eivät olleet tilastollisesti merkittäviä 10 % riskitasolla.

4.2.3 Sammon kartta

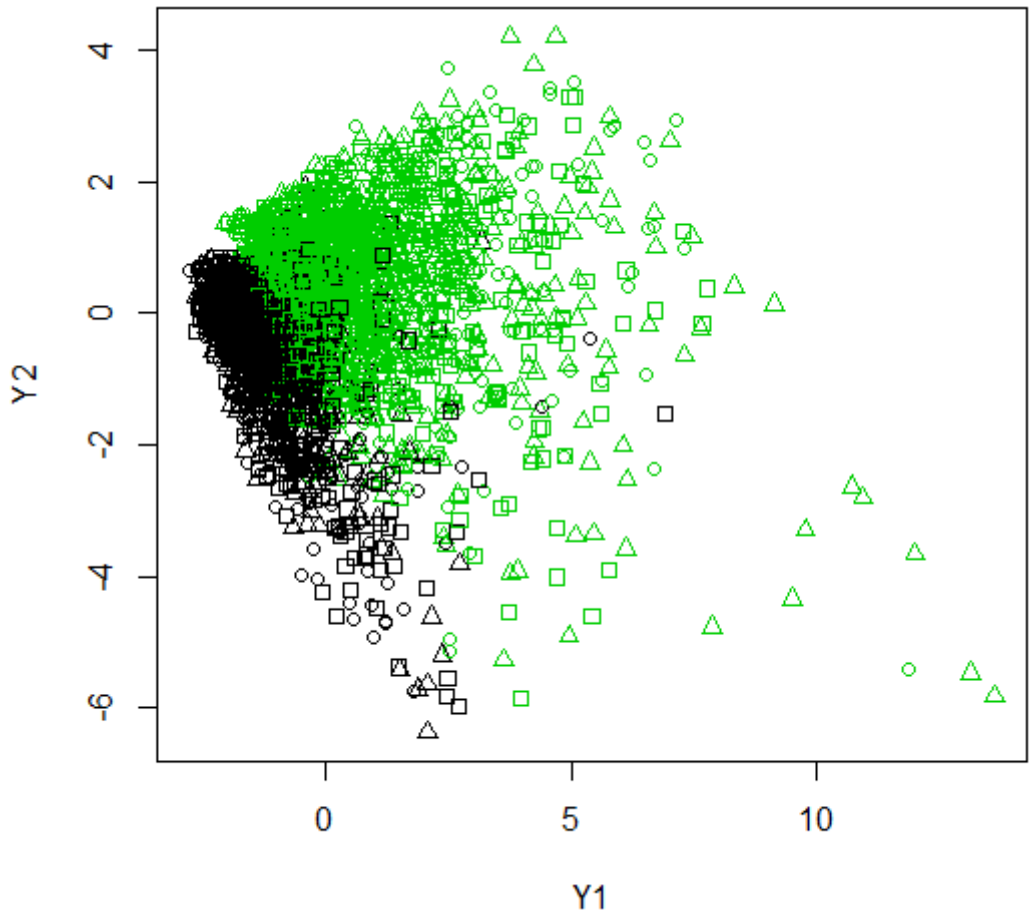
Aiemmin tutkittiin aineistossa havaittavaa vaihtelua siten, että vastaajaryhmät pyrittiin erottelemaan mahdollisimman hyvin. Selvitetään seuraavaksi, millaista vaihtelua muuttujissa on havaittavissa,

kun menetelmän optimoinnissa ei käytetä aiempien analyysien vastemuuttujaa, vaan pelkkiä selittäviä muuttujia. Käytetään tässä Sammon karttaa kaikille selittäville ja tutkitaan sitä, minkälaista vaihtelua tiivistetyssä datassa on.

Iteraation ensimmäisessä askeleessa käytetään tiivistettynä aineistona **Y'** pääkomponenttianalyysin kahta ensimmäistä pääkomponenttia. Käytetään sitten Sammon karttaa tiivistetyn datan optimoimiseen, jolloin saadaan iteraatioiden tuloksena 2-ulotteinen data, jota voidaan kuvata pisteparvena. Tässä ei ole havaittavissa erottelua vastaajaryhmien välillä.

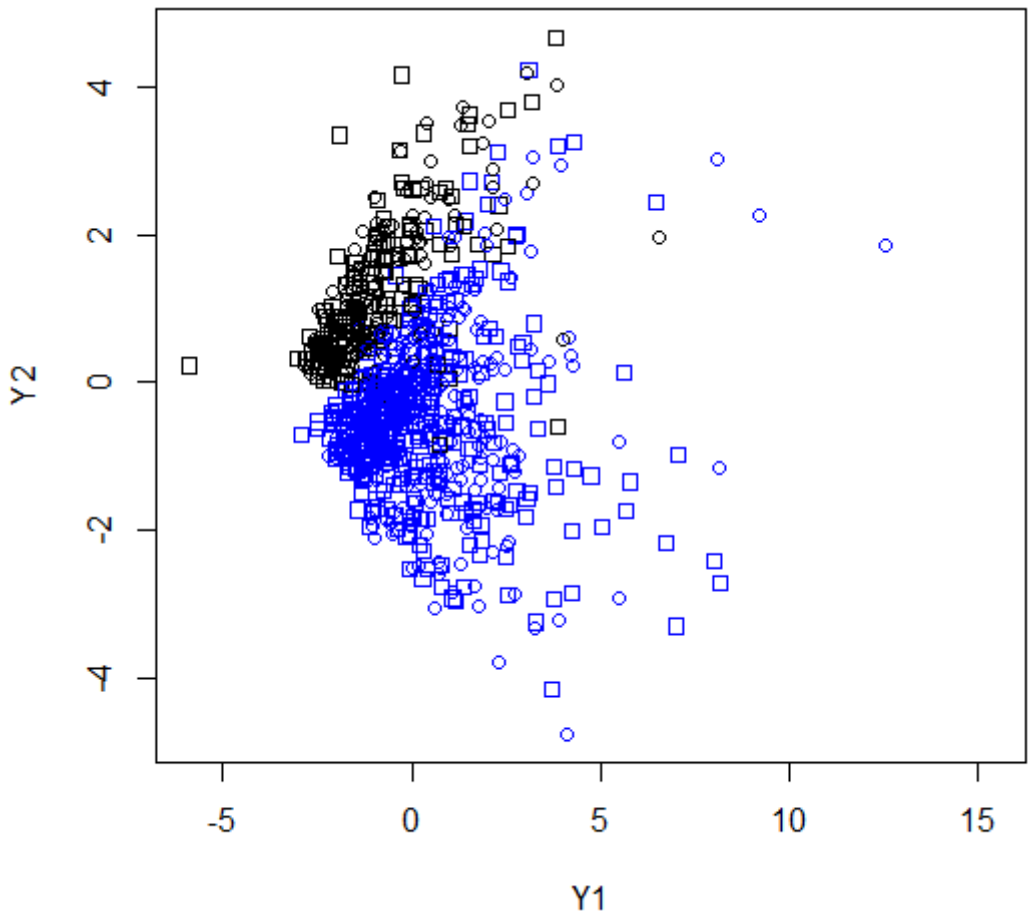
Aloitetaan tarkastelemalla ensimmäisen kyselyn vastauksia. Tällöin analyysiin otetaan kaikki vastaajat, sillä aineistoon ei kuulu ensimmäisen kyselyn yli hypänneitä. Määritetään tälle aineistolle ensin kaksi ensimmäistä pääkomponenttia ja sovelletaan sitten Sammon karttaa käyttämällä pääkomponentteja iteraation ensimmäisessä askeleessa.

Jos vertaillaan muuttujien eri arvojen ryhmien erottuvuutta pisteparvessa, voidaan havaita joidenkin muuttujien arvojen olevan selvästi ryhmittyneitä pisteparvessa. Silmämääräisesti selvimmin erottuu humalahakuinen juominen, joka havaittiin luokitteluanalyysissä tärkeäksi muuttujaksi. Humalahakuisen juomisen kahden ryhmän erottumista Sammon kartassa on havainnollistettu kuviossa 4.4. Kuviossa 2-ulotteisen tiivistetyn datan muuttujat ovat Y1 ja Y2. Humalahakuisen juomisen ryhmää ”ei koskaan” on kuvattu mustalla ja ”vähintään joskus” on vihreällä. Lisäksi vastaajaryhmät on merkitty siten, että kaikkiin vastanneet on neliöitä, viimeisen yli hypänneet ovat ympyröitä ja loput kolmioita.



Kuvio 4.4. Ensimmäisen kyselyn tutkittavien kysymysten Sammon kartta kaksiulotteisessa pisteparvessa, kun humalahakuisen juomisen ryhmät on väritetty eri väreillä sekä eri vastaajaryhmät eri muotoisilla symboleilla.

Tehdään vastaava tarkastelu toisen kyselyn vastauksille. Tässäkään ryhmien välillä ei ole havaittavissa ryhmittymistä pisteparvessa, lisäksi pääasiassa muuttujien arvot eivät näytä erottuvan omina ryhminään. Selvänä poikkeuksena tässä on humalahakuinen juominen, joka näyttää dominoivan myös toisessa kyselyssä. Muuttujan erottumista pisteparvessa on havainnollistettu kuviossa 4.5. jossa humalahakuisen juomisen kaksi ryhmää on korostettu eri väreillä sekä vastaajaryhmät eri muotoisilla merkeillä.



Kuvio 4.5. Toisen kyselyn humalahakuisen juomisen ryhmät sekä vastaajaryhmät Sammon kartan kaksiulotteisessa pisteparvessa. Ryhmää ”ei koskaan” on kuvattu mustilla ja ryhmää ”vähintään joskus” sinisillä merkeillä. Lisäksi kaikkiin vastanneet on merkitty neliöillä ja yhden yli hypänneet palloilla.

Sammon karttojen perusteella vaikuttaisi siltä, että humalahakuinen juominen on merkittävä tekijä tutkittaessa haastateltavien vastausten vaihtelua. Sen sijaan vastaajaryhmien välille ei löydetty tässä selvää erottelua. Huomattavaa myös oli, että aineistoa tiivistäessä lineaariset menetelmät antoivat hyvän tuloksen, sillä Sammon kartassa tehtiin ainoastaan muutamia iteraatiota pääkomponenttien ollessa lähtöarvoina.

5 Yhteenveto

Analyysin ensimmäisessä osassa havaittiin, että haastateltavien ensimmäisen kyselyn vastaukset pääasiassa erosivat toisistaan, kun vertailtiin kolmea vastaajaryhmää keskenään. Muun muassa muuttujissa humalahakuinen juominen ja rikekäyttäytyminen näytti olevan nousevaa trendiä, kun yli hypättyjen kyselyiden määrä kasvoi. Nämä muuttujat havaittiin myös analyysin toisessa osassa tilastollisesti merkittäviksi käytettäessä luokitteluanalyysiä vertailtaessa ryhmiä pareittain.

Aineiston analysoinnin seuraavassa vaiheessa havaittiin useiden nuoren mielenterveyttä ja perhetilannetta kuvaavien muuttujien selittävän tämän vastauskäyttäytymisestä, kun vertailtiin ensimmäisen vaiheen jälkeen lopettaneita kaikkiin vastanneisiin. Merkitsevien muuttujien joukossa olivat muun muassa päihteiden käyttöä ja häiriökäyttäytymistä kuvaavia muuttujia, kun humalahakuinen juominen oli kaikkein tärkein selittäjä. Saatuja malleja voitiin pitää uskottavina, sillä mallit keskimääräisesti ennustivat satunnaisuutta paremmin. Lähes kaikki muuttujat havaittiin merkitseviksi vähintään toisessa luokitteluanalyysitarkastelussa.

Vertaillessa toisessa vaiheessa lopettaneita kaikkiin vastanneisiin havaittiin myös useita vastauskäyttämistä selittäviä muuttujia. Tärkeitä selittäjiä olivat muun muassa häiriökäyttäytyminen, toimettomuus sekä muutamia päihteiden käyttöä kuvaavia muuttujia. Merkitsevistä selittäjistä hieman yli puolet oli jälkimmäisestä kyselystä, joten vaikuttaisi että tuoreemmat vastaukset ennustivat hieman paremmin vastauskäyttäytymistä kuin vanhemmat. Tässäkin analyysissä saadut mallit olivat keskimääräisesti satunnaisuutta merkittävästi parempia.

Analyysissä käytetyistä menetelmistä lineaarinen luokitteluanalyysi vaikutti jonkin verran paremmalta neliölliseen luokitteluanalyysiin verrattuna. Lineaarinen luokitin havaitsi yhteensä enemmän tilastollisesti merkittäviä selittäjiä. Lisäksi lineaariset mallit ennustivat lähes aina neliöllistä luokitinta paremmin otoksessa. Erot eivät kuitenkaan olleet pääasiassa tilastollisesti merkitseviä.

Tutkittaessa haastateltavien vastauksia ilman vastemuuttujana toimivaa vastaajaryhmää havaittiin muuttujilla jonkinlaista ryhmittymistä. Erityisesti aiemmissa analyyseissä tärkeäksi todettu muuttuja humalahakuinen juominen selvästi dominoi vaihtelua. Tämä tuli ilmi sekä ensimmäisessä että toisessa kyselyssä, kun molemmille tehtiin Sammon kartta erikseen.

Lähteet

Achenbach, T. Rescorla, L. (2001), *Manual for the ASEBA school-age forms and profiles,* Burlington VT: University of Vermont, Research Center for Children, Youth and Families.

Blumenthal, J. A. Burg, M. M. Barefoot, J. Williams, R. B. Haney, T. Zimet, G. (1987), *Social support, type A behavior, and coronary artery disease,* Amercian Psychosomatic Society; 252.

Churchill, L. E. Connor, K. M. Davidson, J. R. T. Foa, E. Sherwood, A. Weisler, R. H. (2000), *Psychometric properties of the Social Phobia Inventory (SPIN): New self- rating scale,* Br J Psychiatry; 176, 379–386.

Elisseedd, A. Guyon, I. (2003), *An Introduction to Variable and Feature Selection,* Journal of Machine Learning Research; 1167–1168.

Fröjd, S. Kaltiala-Heino, R. & Marttunen, M. (2010), *Does problem behavior affect attrition from a cohort study on adolescent mental health?* Europian Journal of Public Health, 1–5.

Hollander, M. Wolfe, D. A. (1973), *Nonparametric Statistical Methods,* New York: John Wiley & Sons. 115–120.

Johnson, R. Wichern, D. (2007), *Applied Multivariate Statistical Analysis,* New Jersey: Pearson Prentice Hall, 432, 584–585, 593.

Lindgren, B. W. (1976), *Statistical theory*, London: Collier Macmillan publishers, 41.

Raitasalo, R. (2007), *Mielialakysely Suomen oloihin Beckin lyhyen depressiokyselyn pohjalta kehitetty masennusoireilun ja itsetunnon kysely,* Helsinki: Sosiaali- ja terveysturvan tutkimuksia, 22–24.

Sammon, J. W. (1969), *A nonlinear mapping for data structure analysis,* IEEE Transactions on Computers, 401.

Suominen, S. Koskenvuo, K. Sillanmäki, L. et al (2012), *Non-response in a nationwide follow-up postal survey in Finland: a register-based mortality analysis of respondents and non-respondents of the Health and Social Support (HeSSup) Study*, BMJ Open.

Wilcoxon, W. (1945), *Individual Comparisons by Ranking Methods,* Biometrics Bulletin Vol. 1. No. 6, 80–81.